

## Model-based Kernel Sum Rule

Yu Nishiyama · Motonobu Kanagawa ·  
Arthur Gretton · Kenji Fukumizu

Received: date / Accepted: date

**Abstract** In this study, we enrich the framework of nonparametric kernel Bayesian inference via the flexible incorporation of certain probabilistic models, such as additive Gaussian noise models. Nonparametric inference expressed in terms of kernel means, which is called kernel Bayesian inference, has been studied using basic rules such as the kernel sum rule (KSR), kernel chain rule, kernel product rule, and kernel Bayes' rule (KBR). However, the current framework used for kernel Bayesian inference deals only with nonparametric inference and it cannot allow inference when combined with probabilistic models. In this study, we introduce a novel KSR, called model-based KSR (Mb-KSR), which exploits the knowledge obtained from some probabilistic models of conditional distributions. The incorporation of Mb-KSR into nonparametric kernel Bayesian inference facilitates more flexible kernel Bayesian inference than nonparametric inference. We focus on combinations of Mb-KSR, Non-KSR, and KBR, and we propose a filtering algorithm for state space models, which combines nonparametric learning of the observation process using kernel means and additive Gaussian noise models of the transition dynamics. The idea of the Mb-KSR for additive Gaussian noise models can be extended to more general noise model cases, including a conjugate pair with a positive-definite kernel and a probabilistic model.

---

Yu Nishiyama  
Research Center for Statistical Machine Learning,  
The Institute of Statistical Mathematics, Japan  
E-mail: ynishiyam@gmail.com

Motonobu Kanagawa  
The Graduate University for Advanced Studies, Japan  
E-mail: kanagawa@ism.ac.jp

Arthur Gretton  
Gatsby Computational Neuroscience Unit, CSML, UCL, England  
E-mail: arthur.gretton@gmail.com

Kenji Fukumizu  
The Institute of Statistical Mathematics, Japan  
E-mail: fukumizu@ism.ac.jp

**Keywords** Kernel Bayes’ Rule · Kernel Mean · Kernel Method · Kernel Sum Rule · State Space Model

## 1 Introduction

Kernel methods are powerful tools for developing nonlinear algorithms in machine learning (Schölkopf and Smola, 2002; Steinwart and Christmann, 2008). The basis of kernel methods is to transform data into elements in reproducing kernel Hilbert space (RKHS) and to solve problems in that space by exploiting the reproducing property (kernel trick) of the Hilbert space. A recent trend in kernel methods is to exploit the mean feature element, i.e., the *kernel mean*, of a probability distribution in an RKHS (Smola et al, 2007). The kernel mean  $m_P$  is defined by the expectation of a random feature function  $k(\cdot, X)$  with respect to a probability distribution  $P$  on a measurable space  $(\mathcal{X}, \mathcal{B})$ <sup>1</sup>. The mapping  $P \mapsto m_P$  is called the *kernel mean map*. A positive-definite kernel  $k$  associated with kernel means is called *characteristic* (Fukumizu et al, 2004; Sriperumbudur et al, 2010) if the kernel mean map is injective, i.e., every probability distribution can be distinguished by its kernel mean. The use of characteristic kernels guarantees that a kernel mean  $m_P$  uniquely specifies a probability distribution  $P$  and  $m_P$  may be used in place of  $P$  in probability operations, estimations, and the learning of  $P$ . Many machine learning applications that involve kernel means have been proposed, e.g., density estimations (Smola et al, 2007; Song et al, 2008; McCalman et al, 2013), hypothesis tests (Gretton et al, 2012; Gretton and Györfi, 2010; Fukumizu et al, 2008), Bayesian inference (Song et al, 2009, 2010, 2011; Fukumizu et al, 2013; Song et al, 2013; Kanagawa et al, 2014), classification (Muandet et al, 2012), dimension reduction (Fukumizu and Leng, 2012), and control problems (Grünewälder et al, 2012; Nishiyama et al, 2012; Rawlik et al, 2013; Boots et al, 2013).

In the context of Bayesian inference, the basic probabilistic operations of the sum rule, chain rule, product rule, and Bayes’ rule are kernelized in terms of kernel means and they are referred to as the *kernel sum rule* (KSR), *kernel chain rule* (KCR), *kernel product rule* (KPR), and *kernel Bayes’ rule* (KBR), respectively (Song et al, 2013). Combinations of these rules allow Bayesian inference to be expressed entirely in terms of kernel means. In this study, we refer to Bayesian inference in the kernel mean expression as *kernel Bayesian inference*. Song et al (2011) developed a nonparametric belief propagation method to infer the kernel means of marginals. Song et al (2009) and Fukumizu et al (2013) proposed nonparametric filtering algorithms for state space models, where the transition dynamics and observation process are both represented by kernel means and learned nonparametrically from samples. Control problems (MDP (Grünewälder et al, 2012), POMDP (Nishiyama et al, 2012), path integral control (Rawlik et al, 2013), and predictive state representations (Boots et al, 2013)) have also been developed based on kernel Bayesian inference.

Kernel Bayesian inference has many benefits such as: (i) algorithms can perform nonparametric inference and capture complex conditional relations among variables, (ii) algorithms are similarity-based and the application domains are not

---

<sup>1</sup> Formal definitions are provided in Sect. 2.



**Fig. 1** Kernel mean inference for a three-variable chain.

restricted provided that a positive-definite kernel is defined, and (iii) algorithms can compute expected values without requiring density estimation.

Nonparametric kernel Bayesian inference is powerful, but a weakness of the current framework is that inference algorithms (obtained by combinations of rules, i.e., KSR, KCR, KPR, and KBR) only allow *full nonparametric* inference, i.e., all of the conditional probabilities are learned nonparametrically. Depending on the specific application, the knowledge included in probabilistic models should be combined flexibly with kernel Bayesian inference during full nonparametric inference, e.g., in robot localization problems. In state space models of vision-based robot localization problems, the hidden state is a mobile robot’s position and the observation is an image captured by the robot. The robot’s position should be estimated using a sequence of images. The process of observing the captured image at a specific position in a building is typically complex. By contrast, the transition dynamics of the robot’s position are rather simple and they may be modeled using mathematical models of the physical system, which are encoded as probabilistic models of motion models. Thus, it is desirable to combine nonparametric learning of the observation process using kernel Bayesian inference and probabilistic models of the transition dynamics.

In this study, we propose a novel KSR method, called model-based KSR (Mb-KSR), which exploits the knowledge included in some probabilistic models of conditional distributions. The existing KSR (Song et al, 2013) aims to achieve nonparametric learning of a conditional distribution and we refer to it as nonparametric KSR (Non-KSR). Our Mb-KSR method exploits the tractable conditional kernel means of probabilistic models. The incorporation of Mb-KSR into existing rules, i.e., Non-KSR, KCR, KPR, and KBR, provides a more flexible framework for kernel Bayesian inference, which is combined with probabilistic models. We focus on combinations of Mb-KSR, Non-KSR, and KBR, which are sufficient to develop a filtering algorithm for state space models.

Fig. 1 illustrates inference using a combination of Mb-KSR and Non-KSR. We consider the following three cases. (i) (Left chain) Suppose that both conditional probabilities  $P_{Y|X}$  and  $P_{Z|Y}$  are complex and that the training samples are  $\{(X_i, Y_i)\}_{i=1}^n$  and  $\{(Y_i, Z_i)\}_{i=1}^n$ . (ii) (Middle chain) Suppose that the conditional probability  $P_{Y|X}$  is complex and there is a training sample  $\{(X_i, Y_i)\}_{i=1}^n$ , but  $P_{Z|Y}$  is described simply by a probabilistic model, such as an additive Gaussian noise model. (iii) (Right chain) This is the opposite setting compared with (ii), i.e.,  $P_{Y|X}$  is an additive Gaussian noise model but  $P_{Z|Y}$  is complex and there is a training sample  $\{(Y_i, Z_i)\}_{i=1}^n$ . The inference of  $Z$  given  $X$  in the kernel mean expression is then achieved simply by executing: (i) Non-KSR twice for both  $X$  to  $Y$  and  $Y$  to  $Z$ , (ii) Mb-KSR for  $X$  to  $Y$  and Non-KSR for  $Y$  to  $Z$ , and (iii) Non-KSR for  $X$  to  $Y$  and Mb-KSR for  $Y$  to  $Z$ , respectively. Example 32 illustrates the kernel mean estimators for these cases. By introducing Mb-KSR, a kernel mean estimator is represented by a weighted sum of feature functions but also by a weighted sum of tractable conditional kernel means.

The differences between the Mb-KSR and the Non-KSR are as follows. The Non-KSR contains a regularization parameter. Tuning a regularization parameter is always a burden during computation, e.g., it may require cross-validation. In addition, a regularization parameter with a finite value results in a bias error with Non-KSR. By contrast, the Mb-KSR does not contain a regularization parameter because the smoothness of the regression function is determined by the probabilistic model and its knowledge is reflected. It is natural to expect that if a probabilistic model describes the true conditional distribution well, it is better to learn the probabilistic model first and use it as the Mb-KSR. The Mb-KSR and Non-KSR are compared in experiments reported in Sect. 5.1.

In this study, we focus on additive Gaussian noise model cases for the application of Mb-KSR, but the idea can be extended to more general noise model cases, as described in Appendix A.1. A systematic view is obtained by considering a *conjugate* pair that comprises a probabilistic model and a positive-definite kernel.

By combining probabilistic rules, Non-KSR, KBR, and the proposed Mb-KSR, we develop a filtering algorithm for state space models. In this setting, the observation process is learned nonparametrically by using kernel means and the transition dynamics are given by probabilistic models, such as additive Gaussian noise models.

The main contributions of this study are summarized as follows.

- We propose to split KSR into Non-KSR and Mb-KSR depending on whether the conditional distribution can assume probabilistic models or not. Mb-KSR can be incorporated into existing probabilistic rules, i.e., Non-KSR, Mb-KSR, KCR, KPR, and KBR, thereby yielding a more flexible framework for kernel Bayesian inference when combined with probabilistic models. The Mb-KSR is more accurate and it requires less computation than the Non-KSR if the conditional distribution is described well by probabilistic models, such as additive Gaussian noise models.
- We propose a filtering algorithm, i.e., Algorithm 1, for state space models, which combines nonparametric learning of the observation process using kernel means and probabilistic models of the transition dynamics. The advantages of this algorithm are as follows. The algorithm can handle arbitrary observation domains, e.g., images (as shown in Sect. 5.3), provided that a positive-definite kernel is defined, whereas well-known filtering algorithms such as nonlinear Kalman filters and particle filters restrict the domain to (a subset of) the Euclidean space  $\mathbb{R}^d$ . The proposed algorithm does not assume a specific probabilistic model for the observation process, i.e., the observation process is learned nonparametrically from a training sample, whereas nonlinear Kalman filters and particle filters need to assume a probabilistic model of the observation process.

A related method is described as follows. Kanagawa et al (2014) proposed a Monte-Carlo (MC) sampling method for a filtering algorithm for state space models in the same setting, which combines nonparametric learning of the observation process using kernel means and probabilistic models of the transition dynamics. Because the transition dynamics assume a probabilistic model, the algorithm generates training samples from the transition dynamics and estimates the kernel means using the sample. However, if the transition dynamics are simple, e.g., frequently used additive Gaussian noise models (as considered in the present study),

explicit expressions of the kernel means of probabilistic models can be exploited and MC sampling methods are not necessary. Thus, we propose a filtering algorithm for additive Gaussian transition noise models that does not use MC sampling methods. In general, the difference between the MC method Kanagawa et al (2014) and our proposed method is analogous to the difference between particle filters and Kalman filters, where Kalman filters do not require sampling methods.

The remainder of this paper is structured as follows. The next section provides preliminary details of kernel Bayesian inference, which are used in the later sections that propose the Mb-KSR and the filtering algorithm. Sect. 3 introduces the Mb-KSR. Sect. 4 proposes a filtering algorithm for state space models. Sect. 5 presents the results of ground-truth experiments, which confirm that the proposed Mb-KSR method performs adequately, and we describe a real-world application to robot localization problems. Our conclusions and suggestions for future work are given in Sect. 6.

## 2 Preliminaries: Kernel Bayesian Inference

This section reviews kernel means, KSR, and KBR, which are the components of kernel Bayesian inference (see Song et al (2009); Fukumizu et al (2013); Song et al (2013) for further details). All of the functions considered in this study are real-valued unless stated otherwise explicitly.

### *Positive-definite kernel*

Let  $\mathcal{X}$  be an arbitrary nonempty set. A function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is called symmetric if  $k(x, y) = k(y, x)$  holds for any  $x, y \in \mathcal{X}$ . A symmetric function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is called a *positive-definite kernel* if for  $\forall n \in \mathbb{N}$  and  $\forall x_1, \dots, x_n \in \mathcal{X}$ ,  $n \times n$  matrix  $G = (k(x_i, x_j))_{ij}$ ,  $i, j \in \{1, \dots, n\}$  is positive-semidefinite. The positive-semidefinite matrix  $G$  is called a *Gram matrix*. For each  $x \in \mathcal{X}$ , a function  $k(\cdot, x)$  is called the *feature function* of  $x$ . A positive-definite kernel  $k(x, y)$ ,  $x, y \in \mathbb{R}^m$  is called *shift-invariant* if a function  $\psi : \mathbb{R}^m \rightarrow \mathbb{R}$  exists such that  $k(x, y) = \psi(x - y)$ ,  $x, y \in \mathbb{R}^m$ .  $\psi$  is called a *positive-definite function*.

### *Reproducing kernel Hilbert space (RKHS)*

An RKHS  $\mathcal{H}$  on a nonempty set  $\mathcal{X}$  is a Hilbert space of functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  such that for every  $x \in \mathcal{X}$ , there exists a unique element  $e_x \in \mathcal{H}$  as follows:

$$f(x) = \langle f, e_x \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H} \quad (\text{reproducing property}), \quad (1)$$

where  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  denotes the inner product of the Hilbert space. For any RKHS  $\mathcal{H}$ , the function  $k(x, y) = e_x(y)$ ,  $x, y \in \mathcal{X}$  is a positive-definite kernel. Conversely, for any positive-definite kernel  $k$ , there exists a unique RKHS  $\mathcal{H}$  [Moore-Aronszajn Theorem]. We denote a triplet  $(\mathcal{X}, k, \mathcal{H})$  to indicate that the RKHS  $\mathcal{H}$  is generated by a positive-definite kernel  $k$  on a domain  $\mathcal{X}$ .

### *Kernel mean in a RKHS on a measurable space*

Let  $\mathcal{P}(\mathcal{X})$  be the set of probability distributions on a measurable space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ . Let  $\mathbb{E}_X[f(X)] := \int_{\mathcal{X}} f(x) dP(x)$  denote the expectation of a measurable function  $f : \mathcal{X} \rightarrow \mathbb{R}$  with respect to a probability distribution  $P \in \mathcal{P}(\mathcal{X})$  of a random variable  $X$ . The *kernel mean* of a probability distribution  $P \in \mathcal{P}(\mathcal{X})$  in the RKHS  $\mathcal{H}$  generated by a positive-definite kernel  $k$  is an RKHS element:

$$m_P := \mathbb{E}_X[k(\cdot, X)] \in \mathcal{H}. \quad (2)$$

If  $k$  is shift-invariant such that  $k(x, y) = \psi(x - y)$ ,  $x, y \in \mathbb{R}^m$ , the kernel mean is given by  $m_P = \psi * P$  with the convolution  $*$ . Throughout this study, we assume that a positive-definite kernel is bounded ( $\sup_{x \in \mathcal{X}} k(x, x) < \infty$ ). A bounded kernel guarantees that the kernel mean  $m_P$  is well defined for all  $P \in \mathcal{P}(\mathcal{X})$  (Sriperumbudur et al, 2010). The *expectation property* for the kernel mean  $m_P$  is as follows:

$$\langle m_P, f \rangle_{\mathcal{H}} = \mathbb{E}_X[f(X)], \quad \forall f \in \mathcal{H}. \quad (3)$$

The expectation property is used in applications of kernel Bayesian inference. Typically, the kernel mean is estimated as a weighted sum  $\hat{m}_P = \sum_{i=1}^n w_i k_{\mathcal{X}}(\cdot, \tilde{X}_i)$  with weights  $w = \{w_i\} \in \mathbb{R}^n$  from the data  $\tilde{X}_1, \dots, \tilde{X}_n$ . The expectation of an RKHS function  $f \in \mathcal{H}$  can then be estimated using eq.(3), as follows.

$$\mathbb{E}_X[f(X)] \approx \langle \hat{m}_P, f \rangle_{\mathcal{H}} = \sum_{i=1}^n w_i f(\tilde{X}_i). \quad (4)$$

If  $\hat{m}_P$  is a consistent estimator such that  $\|\hat{m}_P - m_P\|_{\mathcal{H}} \xrightarrow{P} 0$ , then eq. (4) is a consistent estimator. It is not straightforward to show that eq. (4) is consistent for a non-RKHS function  $f \notin \mathcal{H}$ . However, Kanagawa and Fukumizu (2014) proved that eq. (4) is consistent for more general functions in the Besov space under Gaussian RBF kernels. Song et al (2009) and Fukumizu et al (2013) provided consistent estimators for the nonparametric KSR and KBR, as follows.

#### Kernel sum rule (KSR)

Let  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  and  $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$  be measurable spaces, and let  $(X, Y)$  be a random variable on  $\mathcal{X} \times \mathcal{Y}$  with a joint probability distribution  $P_{\mathcal{X} \times \mathcal{Y}}$ . Let  $P_{\mathcal{X}}$  be the marginal distribution of  $P_{\mathcal{X} \times \mathcal{Y}}$  on  $\mathcal{X}$ . For each  $x \in \mathcal{X}$ , let  $P_{\mathcal{Y}|x}$  be the conditional distribution. We denote  $P_{\mathcal{Y}|\mathcal{X}} := \{P_{\mathcal{Y}|x} | x \in \mathcal{X}\}$ . If  $P_{\mathcal{X} \times \mathcal{Y}}$ ,  $P_{\mathcal{X}}$ , and  $P_{\mathcal{Y}|x}$  have densities, we denote them as  $p(x, y)$ ,  $p(x)$ , and  $p(y|x)$ , respectively. Let  $\Pi$  be another probability distribution on  $\mathcal{X}$  and if it exists, then  $\pi(x)$  is its density.  $P_{\mathcal{Y}|\mathcal{X}}$  and  $\Pi$  define a new joint probability distribution  $Q_{\mathcal{X} \times \mathcal{Y}}$  on  $\mathcal{X} \times \mathcal{Y}$  as

$$Q_{\mathcal{X} \times \mathcal{Y}}(A \times B) = \int_A P_{\mathcal{Y}|x}(B) d\Pi(x), \quad A \in \mathcal{B}(\mathcal{X}), \quad B \in \mathcal{B}(\mathcal{Y}). \quad (5)$$

We denote  $Q_{\mathcal{Y}}$  as the marginal of  $Q_{\mathcal{X} \times \mathcal{Y}}$  on  $\mathcal{Y}$  and  $q(y)$  is its density, if it exists. The *sum rule* is a computation of the marginal  $Q_{\mathcal{Y}}$  given an input distribution  $\Pi$ . The sum rule deals with a mapping  $U_{\mathcal{Y}|\mathcal{X}} : \Pi \mapsto Q_{\mathcal{Y}}$  and in density form is  $q(y) = \int_{\mathcal{X}} p(y|x) \pi(x) dx$ .

Let  $(\mathcal{X}, k_{\mathcal{X}}, \mathcal{H}_{\mathcal{X}})$  and  $(\mathcal{Y}, k_{\mathcal{Y}}, \mathcal{H}_{\mathcal{Y}})$ . The KSR is a kernelization of the sum rule, i.e., the computation of the kernel mean  $m_{Q_{\mathcal{Y}}} := \mathbb{E}_{Y \sim Q_{\mathcal{Y}}}[k_{\mathcal{Y}}(\cdot, Y)]$  given an input kernel mean  $m_{\Pi} := \mathbb{E}_{X \sim \Pi}[k_{\mathcal{X}}(\cdot, X)]$ . The KSR deals with a mapping  $\mathcal{U}_{\mathcal{Y}|\mathcal{X}} : m_{\Pi} \mapsto m_{Q_{\mathcal{Y}}}$ . For each  $x \in \mathcal{X}$ , let  $m_{\mathcal{Y}|x}$  be a conditional kernel mean of  $P_{\mathcal{Y}|x}$  defined by  $m_{\mathcal{Y}|x} := \mathbb{E}_{Y|X}[k_{\mathcal{Y}}(\cdot, Y) | X = x] \in \mathcal{H}_{\mathcal{Y}}$ . We write  $m_{\mathcal{Y}|\mathcal{X}} := \{m_{\mathcal{Y}|x} | x \in \mathcal{X}\}$ . Song et al (2009) provided a KSR estimator with nonparametric learning of the conditional kernel mean  $m_{\mathcal{Y}|\mathcal{X}}$  of conditional distribution  $P_{\mathcal{Y}|\mathcal{X}}$ . In this study, we refer to this as the Non-KSR. Let  $\{(X_i, Y_i)\}_{i=1}^n$  be a sample drawn *i.i.d* from  $P_{\mathcal{Y}|\mathcal{X}}$  with an input distribution  $P_{\mathcal{X}}$ . Let  $\hat{m}_{\Pi} = \sum_{i=1}^l \gamma_i k_{\mathcal{X}}(\cdot, \tilde{X}_i)$  be an estimator of the input

kernel mean  $m_{\Pi}$  with a weight vector  $\gamma \in \mathbb{R}^l$  using the data  $(\tilde{X}_1, \dots, \tilde{X}_l)$ . The Non-KSR estimator is given by

$$\text{Non - KSR :} \quad \hat{m}_{Q_Y} = \sum_{j=1}^n w_j k_Y(\cdot, Y_j), \quad w := (G_X + n\varepsilon_n I_n)^{-1} G_{X\tilde{X}} \gamma, \quad (6)$$

where  $G_X = (k_X(X_i, X_j))_{ij} \in \mathbb{R}^{n \times n}$  and  $G_{X\tilde{X}} = (k_X(X_i, \tilde{X}_j))_{ij} \in \mathbb{R}^{n \times l}$ ,  $I_n$  is the  $n \times n$  identity matrix, and  $\varepsilon_n$  is a regularization parameter used to make the inverse operation stable. The Non-KSR deals with the mapping  $\hat{\mathcal{U}}_{Y|\mathcal{X}} : \sum_{i=1}^l \gamma_i k_X(\cdot, \tilde{X}_i) \mapsto \sum_{j=1}^n w_j k_Y(\cdot, Y_j)$ . The consistency  $\|\hat{m}_{Q_Y} - m_{Q_Y}\|_{\mathcal{H}_X} \xrightarrow{P} 0$  of the estimator is proven under  $\|\hat{m}_{\Pi} - m_{\Pi}\|_{\mathcal{H}_X} \xrightarrow{P} 0$  and  $\varepsilon_n \rightarrow 0$  ( $n \rightarrow \infty$ ) with an appropriate rate.

*Kernel Bayes' rule (KBR)*

We use the same notations that are employed above for the KSR. Let  $Q_{\mathcal{X}|y}$  be the conditional distribution of  $Q_{\mathcal{X} \times \mathcal{Y}}$  given  $y \in \mathcal{Y}$ .  $Q_{\mathcal{X}|y}$  is the posterior distribution. Its density form is  $q(x|y) = p(y|x)\pi(x) / \int_{\mathcal{X}} p(y|x)\pi(x)dx$ . The KBR is used to compute the kernel mean  $m_{Q_{\mathcal{X}|y}} := \mathbb{E}_{X \sim Q_{\mathcal{X}|y}}[k_X(\cdot, X)]$  of the posterior  $Q_{\mathcal{X}|y}$ . Fukumizu et al (2013) provided a KBR estimator. Let  $\hat{m}_{\Pi} = \sum_{i=1}^l \gamma_i k_X(\cdot, \tilde{X}_i)$  be an estimator of the prior kernel mean  $m_{\Pi}$  with weights  $\gamma \in \mathbb{R}^l$  and the data  $(\tilde{X}_1, \dots, \tilde{X}_l) \subset \mathcal{X}$ . The KBR estimator is given by

$$\text{KBR :} \quad \hat{m}_{Q_{\mathcal{X}|y}} = \sum_{j=1}^n \tilde{w}_j k_X(\cdot, X_j), \quad \tilde{w} := R_{\mathcal{X}|\mathcal{Y}} \mathbf{k}_Y(y), \quad (7)$$

where  $R_{\mathcal{X}|\mathcal{Y}}$  is an  $n \times n$  matrix such that  $R_{\mathcal{X}|\mathcal{Y}} := D(w)G_Y((D(w)G_Y)^2 + \delta_n I_n)^{-1}D(w)$  and  $\mathbf{k}_Y(y) = (k_Y(y, Y_1), \dots, k_Y(y, Y_n))^{\top}$ . Here,  $D(w)$  is the diagonal matrix of the vector  $w = (G_X + n\varepsilon_n I_n)^{-1} G_{X\tilde{X}} \gamma$ .  $G_Y$  is a Gram matrix such that  $(G_Y)_{ij} = k_Y(Y_i, Y_j)$  and  $\delta_n$  is a regularization parameter. The weight vector  $w$  corresponds to the KSR's weights (6).

### 3 Kernel Bayesian Inference with Probabilistic Models

In this section, we introduce the Mb-KSR method. We explicitly define this method such that combinations of Non-KSR, Mb-KSR, and KBR immediately indicate how to infer the kernel means using different combinations of nonparametric learning and probabilistic models. Sect. 3.1 defines the Mb-KSR. Sect. 3.2 describes inference by combining Non-KSR and Mb-KSR for the chain example shown in Fig. 1. Sect. 3.3 describes a KBR adapted to the Mb-KSR, where the prior kernel mean estimator is slightly different. These results are used to derive the filtering algorithm in Sect. 4.

#### 3.1 Model-based KSR (Mb-KSR)

The Mb-KSR is based on a simple observation with respect to the conditional kernel mean  $m_{Y|\mathcal{X}}$ . Assume that the function  $m_{Y|(\cdot)}(y)$  on  $\mathcal{X}$  with fixed  $y \in \mathcal{Y}$  is

included in the RKHS  $\mathcal{H}_{\mathcal{X}}$ . Then,  $m_{Q_{\mathcal{Y}}}(y)$  is given in relation to  $m_{\Pi}$  as

$$\begin{aligned} m_{Q_{\mathcal{Y}}}(y) &= \mathbb{E}_{X \sim \Pi} \mathbb{E}_{Y|X} [k_{\mathcal{Y}}(y, Y)|X] = \mathbb{E}_{X \sim \Pi} [m_{\mathcal{Y}|X}(y)] \\ &= \mathbb{E}_{X \sim \Pi} [\langle k_{\mathcal{X}}(\cdot, X), m_{\mathcal{Y}|\cdot}(y) \rangle_{\mathcal{H}_{\mathcal{X}}}] = \langle m_{\Pi}, m_{\mathcal{Y}|\cdot}(y) \rangle_{\mathcal{H}_{\mathcal{X}}}. \end{aligned}$$

Thus,  $m_{Q_{\mathcal{Y}}}(y)$  is obtained by the inner product of  $m_{\Pi}$  and  $m_{\mathcal{Y}|\cdot}(y)$ .  $m_{Q_{\mathcal{Y}}}(y)$  can be estimated using the input kernel mean estimator  $\hat{m}_{\Pi} = \sum_{i=1}^l \gamma_i k_{\mathcal{X}}(\cdot, \tilde{X}_i)$  as

$$m_{Q_{\mathcal{Y}}}(y) \approx \langle \hat{m}_{\Pi}, m_{\mathcal{Y}|\cdot}(y) \rangle_{\mathcal{H}_{\mathcal{X}}} = \sum_{i=1}^l \gamma_i m_{\mathcal{Y}|\tilde{X}_i}(y). \quad (8)$$

We then define the following estimator:

$$\text{Mb-KSR} : \quad \hat{m}_{Q_{\mathcal{Y}}} := \sum_{i=1}^l \gamma_i m_{\mathcal{Y}|\tilde{X}_i}. \quad (9)$$

The Mb-KSR deals with the mapping  $\bar{\mathcal{U}}_{\mathcal{Y}|\mathcal{X}} : \sum_{i=1}^l \gamma_i k_{\mathcal{X}}(\cdot, \tilde{X}_i) \mapsto \sum_{i=1}^l \gamma_i m_{\mathcal{Y}|\tilde{X}_i}$ . Since  $\{m_{\mathcal{Y}|\tilde{X}_i}\}_{i=1}^l$  are elements in RKHS  $\mathcal{H}_{\mathcal{Y}}$ , a linear combination of them,  $\hat{m}_{Q_{\mathcal{Y}}}$ , is also an element in RKHS  $\mathcal{H}_{\mathcal{Y}}$ . In Proposition A10 (Appendix A.2), we prove that if  $\hat{m}_{\Pi}$  is a consistent estimator of  $\|\hat{m}_{\Pi} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}} \xrightarrow{P} 0$ , then eq. (9) is a consistent estimator of  $\|\hat{m}_{Q_{\mathcal{Y}}} - m_{Q_{\mathcal{Y}}}\|_{\mathcal{H}_{\mathcal{X}}} \xrightarrow{P} 0$ . The consistency rate of  $\hat{m}_{Q_{\mathcal{Y}}}$  is the same as that of the input kernel mean  $\hat{m}_{\Pi}$ .

In the Mb-KSR, we consider the conditional distribution  $P_{\mathcal{Y}|x}$  as a probabilistic model, with  $m_{\mathcal{Y}|x}$  its conditional kernel mean. If  $m_{\mathcal{Y}|x}(y)$  is computed in an efficient manner, then eq. (8) is computed efficiently. For example, if  $P_{\mathcal{Y}|x}$  is an additive Gaussian noise model and  $k$  is a Gaussian kernel, then  $m_{\mathcal{Y}|x}$  has an explicit expression and  $m_{\mathcal{Y}|x}(y)$  is computed efficiently for every  $y \in \mathcal{Y}$ . In the following, we consider the Mb-KSR of an additive Gaussian noise model. Thus, we focus on the Gaussian case in the present study.

Let  $d_G(x|\mu, R) := \frac{1}{\sqrt{|2\pi R|}} \exp(-\frac{1}{2}(x - \mu)^{\top} R^{-1}(x - \mu))$ ,  $x \in \mathbb{R}^m$  denote the  $m$ -dimensional Gaussian density function with mean vector  $\mu$  and covariance matrix  $R$ . We denote a Gaussian random vector  $X$  on  $\mathbb{R}^m$  as  $X \sim N(\mu, R)$ . Let  $k_R(x_1, x_2) := d_G(x_1 - x_2|0, R)$ ,  $x_1, x_2 \in \mathbb{R}^m$  denote a Gaussian kernel<sup>2</sup> and  $(\mathbb{R}^m, k_R, \mathcal{H}_R^{(G)})$  be the corresponding RKHS.

**Example 31 (additive Gaussian noise model)** Let  $\mathcal{Y} = \mathbb{R}^m$ . Assume that  $\{P_{\mathcal{Y}|x}|x \in \mathcal{X}\}$  is an additive Gaussian noise model  $y = f(x) + \epsilon$  with a function  $f : \mathcal{X} \rightarrow \mathbb{R}^m$  and a Gaussian noise  $\epsilon \sim N(0, \Sigma)$ . The conditional kernel mean  $m_{\mathcal{Y}|x}$  of the additive Gaussian noise model in a Gaussian RKHS  $\mathcal{H}_R^{(G)}$  is given by

$$m_{\mathcal{Y}|x} = d_G(\cdot|f(x), \Sigma + R) \in \mathcal{H}_R^{(G)}. \quad (10)$$

<sup>2</sup> For convenience, we consider unnormalized positive definite kernels in the present study. The kernel mean  $m_P$  in eq. (2) is the same up to a constant multiplication even if the normalized positive-definite kernel  $k_R(x_1, x_2) := \exp(-\frac{1}{2}(x_1 - x_2)^{\top} R^{-1}(x_1 - x_2))$ ,  $x_1, x_2 \in \mathbb{R}^m$  is used.



*Proof* The additive Gaussian noise model has the conditional density  $p(y|x) = d_G(y|f(x), \Sigma)$ . For each  $x \in \mathcal{X}$ , the conditional kernel mean  $m_{\mathcal{Y}|x}$  of the Gaussian density  $d_G(y|f(x), \Sigma)$  in Gaussian RKHS  $\mathcal{H}_R^{(G)}$  is given by a convolution  $m_{\mathcal{Y}|x} = d_G(\cdot|0, R) * d_G(\cdot|f(x), \Sigma)$ , where  $d_G(\cdot|0, R)$  is a positive-definite function of the Gaussian kernel  $k_R$ . In general, the convolution of two Gaussians,  $d_G(\cdot|\mu_1, \Sigma_1) * d_G(\cdot|\mu_2, \Sigma_2)$ , is also a Gaussian  $d_G(\cdot|\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$ , so that  $m_{\mathcal{Y}|x} = d_G(\cdot|f(x), \Sigma + R)$ , which completes the proof.

By substituting the explicit expression (10) into eq. (9), we obtain the Mb-KSR used for computing  $\hat{m}_{Q_{\mathcal{Y}}}$  in the case of the additive Gaussian noise model. For each  $y \in \mathcal{Y}$ , the evaluation  $\hat{m}_{Q_{\mathcal{Y}}}(y)$  of the output kernel mean often needs to be computed in applications. This requires the evaluation of the Gaussian density function (10) in eq.(8). Thus, we obtain the Mb-KSR of an additive Gaussian noise model.

In this study, we focus on additive Gaussian noise models, which are used frequently, but the same idea can be applied to other probabilistic models provided that for each  $y \in \mathcal{Y}$ ,  $m_{\mathcal{Y}|x}(y)$  can be computed. A systematic approach to such probabilistic models is to focus on the *conjugate* pair of a probabilistic model and a positive-definite kernel for the kernel mean (Nishiyama and Fukumizu, 2014). A probabilistic model  $p$  and a positive-definite kernel  $k$  are conjugate if  $p$  and its kernel mean  $m_p$  have the same density form. The pair of an additive Gaussian noise model  $p(y|x)$  and a Gaussian kernel  $k$  is an example, where its kernel mean is also a Gaussian (10). For other examples, see Appendix A.1.

The Non-KSR and Mb-KSR estimators differ as follows. The Non-KSR (6) estimates the output kernel mean  $m_{Q_{\mathcal{Y}}}$  by a linear combination of feature functions  $\{k_{\mathcal{Y}}(\cdot, Y_i) | i = 1, \dots, n\}$  on  $\mathcal{Y}$  with different weights  $w \in \mathbb{R}^n$ , whereas the Mb-KSR (9) estimates  $m_{Q_{\mathcal{Y}}}$  by a linear combination of conditional kernel means  $\{m_{\mathcal{Y}|\tilde{X}_i} | i = 1, \dots, l\}$  with input weights  $\gamma \in \mathbb{R}^l$ , where the evaluation of the conditional kernel means is computationally tractable. If a conjugate pair is used, the conditional kernel means are simply the same functions as feature functions but with increased parameter values. In the additive Gaussian noise model case, the Mb-KSR is given by a linear combination of Gaussian feature functions  $\{d_G(\cdot|f(\tilde{X}_i), \Sigma + R)\}_{i=1}^l$  where the variance is increased by  $\Sigma$ . The Non-KSR estimator uses the regularization parameter  $\epsilon_n$  to determine the smoothness, whereas the Mb-KSR does not and it does not need to be tuned. This is because the smoothness is determined by the probabilistic model and it reflects the knowledge included in the probabilistic model.

### 3.2 Combining the Non-KSR and Mb-KSR

Given the two types of kernel sum rule, Non-KSR and Mb-KSR, we can infer kernel means by combining them. We perform inference for the chain case shown in Fig. 1.

Let  $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ ,  $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$  and  $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$  be measurable spaces. Let  $(X, Y, Z)$  be a random variable on  $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$  with probability distribution  $Q_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}$ . Suppose that  $Q_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}$  is factored into a chain, as in Fig. 1, where it comprises an input distribution  $\Pi$  on  $\mathcal{X}$  and conditional distributions  $P_{\mathcal{Y}|\mathcal{X}}$  and  $P_{\mathcal{Z}|\mathcal{Y}}$ . Let  $Q_{\mathcal{Y}}$  and  $Q_{\mathcal{Z}}$

denote the marginals of  $Q_{\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}}$  on  $\mathcal{Y}$  and  $\mathcal{Z}$ , respectively. Define three RKHSs:  $(\mathcal{X}, k_{\mathcal{X}}, \mathcal{H}_{\mathcal{X}})$ ,  $(\mathcal{Y}, k_{\mathcal{Y}}, \mathcal{H}_{\mathcal{Y}})$ , and  $(\mathcal{Z}, k_{\mathcal{Z}}, \mathcal{H}_{\mathcal{Z}})$ .

**Example 32 (chain)** Let  $\hat{m}_{\Pi} = \sum_{i=1}^l \gamma_i k_{\mathcal{X}}(\cdot, \tilde{X}_i)$  be an estimator of the input kernel mean  $m_{\Pi}$  of a probability distribution  $\Pi$ .

(i) Fig. 1 (Left). This case corresponds to the preliminary results of (Song et al, 2009; Fukumizu et al, 2013; Song et al, 2013) and it is obtained by applying the Non-KSR twice.

(ii) Fig. 1 (Middle). Let  $\{(X_i, Y_i)\}_{i=1}^n$  be the training data for a conditional distribution  $P_{\mathcal{Y}|\mathcal{X}}$ , which are generated from an input distribution  $P_{\mathcal{X}}$ . Suppose that a probabilistic model is given for  $P_{\mathcal{Z}|\mathcal{Y}}$ . The output kernel mean  $m_{Q_{\mathcal{Z}}}$  is then estimated by  $\hat{m}_{Q_{\mathcal{Z}}} = \bar{U}_{\mathcal{Z}|\mathcal{Y}} \hat{U}_{\mathcal{Y}|\mathcal{X}} \hat{m}_{\Pi}$ , where  $\hat{U}_{\mathcal{Y}|\mathcal{X}}$  denotes the Non-KSR operation and  $\bar{U}_{\mathcal{Z}|\mathcal{Y}}$  denotes the Mb-KSR operation, i.e.,

$$\hat{m}_{Q_{\mathcal{Z}}} = \sum_{i=1}^n w_i m_{\mathcal{Z}|Y_i}, \quad w = (G_X + n\epsilon_n I_n)^{-1} G_{X\tilde{X}} \gamma, \quad (11)$$

where  $G_X$  and  $G_{X\tilde{X}}$  are kernel matrices  $G_X = [k_{\mathcal{X}}(X_i, X_j)] \in \mathbb{R}^{n \times n}$  and  $G_{X\tilde{X}} = [k_{\mathcal{X}}(X_i, \tilde{X}_j)] \in \mathbb{R}^{n \times l}$ ,  $I_n$  is the  $n \times n$  identity matrix, and  $\epsilon_n$  is a regularization parameter. The RKHS elements  $\{m_{\mathcal{Z}|Y_i}\}_{i=1}^n \subset \mathcal{H}_{\mathcal{Z}}$  are the conditional kernel means of the probabilistic model  $\{P_{\mathcal{Z}|Y_i}\}_{i=1}^n$  at data points  $\{Y_i\}_{i=1}^n$ .

(iii) Fig. 1 (Right). This is the opposite of setting (ii). Suppose that a probabilistic model is given for  $P_{\mathcal{Y}|\mathcal{X}}$ . Let  $\{(Y_i, Z_i)\}_{i=1}^n$  be the training data for a conditional distribution  $P_{\mathcal{Z}|\mathcal{Y}}$  generated by an input distribution  $P_{\mathcal{Y}}$ . The output kernel mean  $m_{Q_{\mathcal{Z}}}$  is then estimated by  $\hat{m}_{Q_{\mathcal{Z}}} = \hat{U}_{\mathcal{Z}|\mathcal{Y}} \bar{U}_{\mathcal{Y}|\mathcal{X}} \hat{m}_{\Pi}$ , where  $\bar{U}_{\mathcal{Y}|\mathcal{X}}$  denotes the Mb-KSR operation and  $\hat{U}_{\mathcal{Z}|\mathcal{Y}}$  denotes the Non-KSR operation, i.e.,

$$\hat{m}_{Q_{\mathcal{Z}}} = \sum_{i=1}^n w_i k_{\mathcal{Z}}(\cdot, Z_i), \quad w = (G_Y + n\epsilon_n I_n)^{-1} G_{Y|\tilde{X}} \gamma, \quad (12)$$

where  $G_Y$  is the Gram matrix  $G_Y = [k_{\mathcal{Y}}(Y_i, Y_j)]$ ,  $I_n$  is the  $n \times n$  identity matrix, and  $\epsilon_n$  is a regularization parameter.  $G_{Y|\tilde{X}}$  is an  $n \times l$  matrix such that  $(G_{Y|\tilde{X}})_{ij} = m_{Y|\tilde{X}_j}(Y_i)$ , which is obtained by computing the conditional kernel mean  $m_{Y|\tilde{X}_i}$  of a probabilistic model  $P_{\mathcal{Y}|\tilde{X}_i}$  at evaluation points  $\{Y_i\}_{i=1}^n$ .

*Proof* Eqs (11) and (12) are obtained by applying eqs. (6) and (9). Eq. (11) is obtained by

$$\hat{m}_{Q_{\mathcal{Z}}} = \bar{U}_{\mathcal{Z}|\mathcal{Y}} \hat{U}_{\mathcal{Y}|\mathcal{X}} \hat{m}_{\Pi} = \bar{U}_{\mathcal{Z}|\mathcal{Y}} \sum_{i=1}^n w_i k_{\mathcal{Y}}(\cdot, Y_j) = \sum_{i=1}^n w_i m_{\mathcal{Z}|Y_i},$$

where  $w = (G_X + n\epsilon_n I_n)^{-1} G_{X\tilde{X}} \gamma$ .

Eq. (12) is obtained by

$$\begin{aligned}
\hat{m}_{Q_Z} &= \hat{\mathcal{U}}_{Z|Y} \bar{\mathcal{U}}_{Y|X} \hat{m}_\Pi = \hat{\mathcal{U}}_{Z|Y} \sum_{j=1}^l \gamma_j m_{Y|\tilde{X}_j} = \mathbf{k}_Z^\top (G_Y + n\epsilon I_n)^{-1} \mathbf{k}_Y \sum_{j=1}^l \gamma_j m_{Y|\tilde{X}_j} \\
&= \mathbf{k}_Z^\top (G_Y + n\epsilon I_n)^{-1} \begin{pmatrix} \langle k_Y(\cdot, Y_1), \sum_{j=1}^l \gamma_j m_{Y|\tilde{X}_j} \rangle_{\mathcal{H}_Y} \\ \vdots \\ \langle k_Y(\cdot, Y_n), \sum_{j=1}^l \gamma_j m_{Y|\tilde{X}_j} \rangle_{\mathcal{H}_Y} \end{pmatrix} \\
&= \mathbf{k}_Z^\top (G_Y + n\epsilon I_n)^{-1} \begin{pmatrix} \sum_{j=1}^l \gamma_j m_{Y|\tilde{X}_j}(Y_1) \\ \vdots \\ \sum_{j=1}^l \gamma_j m_{Y|\tilde{X}_j}(Y_n) \end{pmatrix} = \sum_{i=1}^n w_i k_Z(\cdot, Z_i),
\end{aligned}$$

where  $w = (G_Y + n\epsilon I_n)^{-1} G_Y \bar{\mathcal{U}}_{Y|X} \gamma$ . For the third equality, we use  $\hat{\mathcal{U}}_{Z|Y} = \mathbf{k}_Z^\top (G_Y + n\epsilon I_n)^{-1} \mathbf{k}_Y$  with notations  $\mathbf{k}_Y = (k_Y(\cdot, Y_1), \dots, k_Y(\cdot, Y_n))^\top$  and  $\mathbf{k}_Z^\top = (k_Z(\cdot, Z_1), \dots, k_Z(\cdot, Z_n))$ , as described previously (Song et al, 2009; Fukumizu et al, 2013; Song et al, 2013).

The consistency of the two estimators, (11) and (12), is addressed in Example A11 (Appendix A.2). Since the rate of the output kernel mean of the Mb-KSR is equal to the rate of the input kernel mean, the two estimators, (11) and (12), have the same consistency rate.

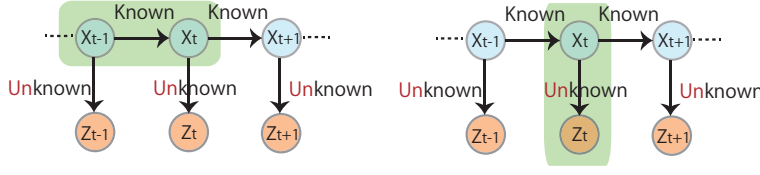
### 3.3 KBR

As described in Sect. 3.1, when the Mb-KSR is considered, a kernel mean estimator is expressed as the weighted sum form of the feature functions  $\hat{m}_P = \sum_{i=1}^l \gamma_i k_{\mathcal{X}}(\cdot, \tilde{X}_i)$  but also the conditional kernel means (9). The KBR (7) when  $\hat{m}_\Pi$  is expressed as a weighted sum of the conditional kernel means can also be obtained in a slightly modified manner, as follows.

Let  $\hat{m}_\Pi = \sum_{j=1}^l \gamma_j m_j$  be a prior kernel mean estimator that uses RKHS elements  $\{m_j\}_{j=1}^l \subset \mathcal{H}_{\mathcal{X}}$ . The KBR is given by

$$\mathbf{KBR} : \quad \hat{m}_{Q_{\mathcal{X}|Y}} = \sum_{j=1}^n \tilde{w}_j k_{\mathcal{X}}(\cdot, X_j), \quad \tilde{w} := R_{\mathcal{X}|Y} \mathbf{k}_Y(y), \quad (13)$$

where  $R_{\mathcal{X}|Y}$  is an  $n \times n$  matrix such that  $R_{\mathcal{X}|Y} := D(w) G_Y ((D(w) G_Y)^2 + \delta_n I_n)^{-1} D(w)$  with a new weight vector  $w = (G_X + n\epsilon_n I_n)^{-1} M \gamma$ . Here,  $M$  is an  $n \times l$  matrix such that  $M_{ij} = m_j(X_i)$ . The new weight vector  $w$  is a result of the Mb-KSR given in Example 32 (iii).



**Fig. 2** State space models. Left: the prediction step is the kernel sum rule (KSR). Right: the filtering step is the kernel Bayes' rule (KBR).

#### 4 Filtering for State Space Models

Using the components described in Sections 2 and 3, we develop a filtering algorithm for state space models (Fig. 2). We consider the following setting for the state space models:

- Let  $x \in \mathcal{X} = \mathbb{R}^m$  be a hidden state. The transition dynamics comprise an additive Gaussian noise model  $x_{t+1} = f(x_t) + \varsigma_t$ ,  $\varsigma_t \sim N(\mathbf{0}, \Sigma)$  with a linear/nonlinear function  $f: \mathbb{R}^m \rightarrow \mathbb{R}^m$ .
- Let  $z$  be an observation in an arbitrary domain  $\mathcal{Z}$ . The observation process is given by a conditional distribution  $P_{\mathcal{Z}|\mathcal{X}}$ . The density is  $p(z|x)$  if it exists. We learn  $P_{\mathcal{Z}|\mathcal{X}}$  nonparametrically from the data  $\{(X_i, Z_i)\}_{i=1}^n$ .

The filtering task comprises the sequential estimation of the hidden state  $x_t$  from a sequence of observations  $z_{1:t} := (z_1, \dots, z_t)$  for each time  $t$ . Due to the Markov property of the state space model, the current state  $x_t$  is estimated using the distribution (belief) of the previous hidden state  $x_{t-1}$  and a current observation  $z_t$ . Fig. 2 shows one step of the filtering algorithm at time  $t$ . The algorithm comprises two steps: the *prediction step* and *filtering step*. Suppose that  $q(x_{t-1}|z_{1:t-1})$  is a distribution of  $x_{t-1}$  given observations  $z_{1:t-1}$ . In the prediction step,  $x_t$  is predicted as  $p(x_t|z_{1:t-1}) = \int p(x_t|x_{t-1})q(x_{t-1}|z_{1:t-1})dx_{t-1}$  without observation  $z_t$ . In the filtering step, the predictive distribution is improved to  $q(x_t|z_{1:t}) = p(z_t|x_t)p(x_t|z_{1:t-1}) / \int_{\mathcal{X}} p(z_t|x)p(x_t|z_{1:t-1})dx_t$  using a prior  $p(x_t|z_{1:t-1})$  and a likelihood  $p(z_t|x_t)$  with a new observation  $z_t$ . The prediction and filtering steps are employed sequentially for each time  $t$ . The prediction step corresponds to the sum rule and the filtering step corresponds to Bayes' rule.

Filtering algorithms that use kernel means can be obtained by applying KSR and KBR. Let  $(\mathcal{X}, k_{\mathcal{X}}, \mathcal{H}_{\mathcal{X}})$  and  $(\mathcal{Z}, k_{\mathcal{Z}}, \mathcal{H}_{\mathcal{Z}})$ . Let  $m_{X_{t-1}|z_{1:t-1}}$ ,  $m_{X_t|z_{1:t-1}}$ , and  $m_{X_t|z_{1:t}}$  be kernel means of  $q(x_{t-1}|z_{1:t-1})$ ,  $p(x_t|z_{1:t-1})$ , and  $q(x_t|z_{1:t})$  in  $\mathcal{H}_{\mathcal{X}}$ , respectively. Suppose that we have an estimator for  $m_{X_{t-1}|z_{1:t-1}}$  as

$$\hat{m}_{X_{t-1}|z_{1:t-1}} = \sum_{i=1}^n [\alpha_{X_{t-1}|z_{1:t-1}}]_i k_{\mathcal{X}}(\cdot, X_i)$$

with weights  $\alpha_{X_{t-1}|z_{1:t-1}} \in \mathbb{R}^n$  and data  $(X_1, \dots, X_n)$ . In the prediction step,  $\hat{m}_{X_t|z_{1:t-1}} = \mathcal{U}_{\mathcal{X}'|\mathcal{X}} \hat{m}_{X_{t-1}|z_{1:t-1}}$ , where  $\mathcal{U}_{\mathcal{X}'|\mathcal{X}}$  is the Mb-KSR operation. From eq. (9), this is given by

$$\hat{m}_{X_t|z_{1:t-1}} = \sum_{i=1}^n [\alpha_{X_{t-1}|z_{1:t-1}}]_i m_{\mathcal{X}'|X_i},$$

---

**Algorithm 1** KBR-Filter (Transition: probabilistic model, Observation: nonparametric)

---

**Initial Belief:** prior kernel mean  $m_{X_1}$ .  
**Observation:**  $z_1 \in \mathcal{Z}$ .  
**Initial Filtering:**  $\alpha_{X_1|z_1} \leftarrow$  KBR Algorithm with prior  $m_{X_1}$ .  
**for**  $t = 2 : T$  **do**  
    **Weight Computation 1:**  $\beta_{Z_t|z_{1:t-1}} = (G_X + \epsilon n I_n)^{-1} G_{X'|X} \alpha_{X_{t-1}|z_{1:t-1}}$ .  
    **Observation:**  $z_t \in \mathcal{Z}$ .  
    **Weight Computation 2:**  $\alpha_{X_t|z_{1:t}} = R_{\mathcal{X}|\mathcal{Z}}(\beta_{Z_t|z_{1:t-1}}) \mathbf{k}_Z(z_t)$ .  
**end for**

---

where  $m_{\mathcal{X}'|X_i}$  is the conditional kernel mean of the transition dynamics (additive Gaussian noise model) given a data point  $X_i$ . In the filtering step,  $m_{X_t|z_{1:t}}$  is obtained by the KBR (13) with the prior kernel mean  $\hat{m}_{X_t|z_{1:t-1}}$ . From eq. (13), the posterior kernel mean is estimated by

$$\hat{m}_{X_t|z_{1:t}} = \sum_{i=1}^n [\alpha_{X_t|z_{1:t}}]_i k_{\mathcal{X}}(\cdot, X_i), \quad \alpha_{X_t|z_{1:t}} = R_{\mathcal{X}|\mathcal{Z}}(\beta_{Z_t|z_{1:t-1}}) \mathbf{k}_Z(z_t),$$

where  $R_{\mathcal{X}|\mathcal{Z}}(\beta_{Z_t|z_{1:t-1}})$  is an  $n \times n$  matrix that depends on  $\beta_{Z_t|z_{1:t-1}}$  as

$$R_{\mathcal{X}|\mathcal{Z}}(\beta_{Z_t|z_{1:t-1}}) := D(\beta_{Z_t|z_{1:t-1}}) G_Z ((D(\beta_{Z_t|z_{1:t-1}}) G_Z)^2 + \delta_n I_n)^{-1} D(\beta_{Z_t|z_{1:t-1}}),$$

$$\beta_{Z_t|z_{1:t-1}} = (G_X + n \epsilon_n I_n)^{-1} G_{X'|X} \alpha_{X_{t-1}|z_{1:t-1}},$$

where  $G_{X'|X}$  is given by  $(G_{X'|X})_{ij} = m_{\mathcal{X}'|X_j}(X_i)$ . This corresponds to eq. (12) in Example 32.

As a result, the filtering algorithm is summarized in Algorithm 1. The next section presents the results of numerical experiments. As described in eq. (4), the expectation  $\mathbb{E}_{X \sim Q_{\mathcal{X}|z_{1:t}}}[f(X)]$  of any RKHS function  $f \in \mathcal{H}_{\mathcal{X}}$  with respect to the posterior distribution  $Q_{\mathcal{X}|z_{1:t}}$  can be estimated as  $\mathbb{E}_{X \sim Q_{\mathcal{X}|z_{1:t}}}[f(X)] \approx \sum_{i=1}^n [\alpha_{X_t|z_{1:t}}]_i f(X_i)$ . The point estimation of  $\hat{x}_t$  from a given kernel mean estimator  $\hat{m}_{X_t|z_{1:t}}$  may be considered by solving a preimage problem (Song et al, 2009; Fukumizu et al, 2011),

$$\hat{x}_t := \arg \min_x \|k_{\mathcal{X}}(\cdot, x) - \hat{m}_{X_t|z_{1:t}}\|_{\mathcal{H}_{\mathcal{X}}}^2, \quad (14)$$

where the minimization algorithm results in a simple iteration in the Gaussian kernel case (Mika et al, 1999).

Comparisons with other typical filtering algorithms are described in the following. Well-known algorithms such as the Kalman filter, extended Kalman filter, unscented Kalman filter, and particle filters restrict the observation domain  $\mathcal{Z}$  to (a subset of) the Euclidean space  $\mathbb{R}^d$ , and they require the setting of a probabilistic model of the observation process  $P_{\mathcal{Z}|\mathcal{X}}$ , e.g., to compute the likelihood in particle filters. By contrast, Algorithm 1 permits an arbitrary observation domain  $\mathcal{Z}$  and does not assume a specific probabilistic model for the observation process.

## 5 Experiments

In this section, we present the results of numerical experiments obtained using the Mb-KSR and the filtering algorithm, Algorithm 1, for state space models. Sect. 5.1 describes simple ground-truth experiments, which validate the concept of the Mb-KSR and they illustrate the differences compared with the Non-KSR. Sect. 5.2 presents filtering results for synthetic nonlinear state space models, which demonstrate the superior performance of the proposed Algorithm 1 compared with the existing full nonparametric kernel Bayes filter (Fukumizu et al, 2013, Sect. 4.3) due to the incorporation of a probabilistic model. Sect. 5.3 describes the application of our proposed method to real-world vision-based robot localization problems.

### 5.1 Ground-truth Experiment

We validate the Mb-KSR concept and determine how it differs from the Non-KSR. We consider a case where the true output kernel mean  $m_{Q_Y}$  described in Sect. 3.1 has an analytical solution. Next, we evaluate the accuracy  $\|m_{Q_Y} - \hat{m}_{Q_Y}\|_{\mathcal{H}_Y}$  of the estimators in the RKHS norm.

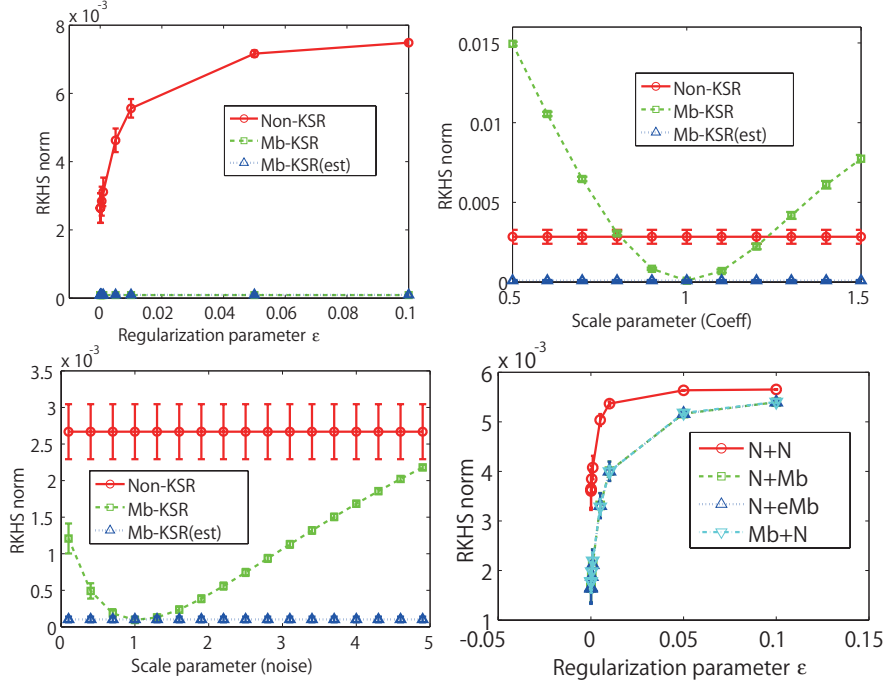
Let  $\mathcal{X} = \mathcal{Y} = \mathbb{R}^m$  and  $P_{Y|\mathcal{X}}$  be a linear Gaussian model  $y = Ax + \epsilon$  with matrix  $A \in \mathbb{R}^{m \times m}$  and Gaussian noise  $\epsilon \sim N(\mathbf{0}, \Sigma)$ . If the input distribution  $\Pi$  is a Gaussian mixture with density  $\pi(x) = \sum_{i=1}^L \xi_i d_G(x; \mu_i, W_i)$ , then  $Q_Y$  is also a Gaussian mixture with density  $q(y) = \int p(y|x)\pi(x)dx = \sum_{i=1}^L \xi_i d_G(y; A\mu_i, \Sigma + AW_iA^\top)$ . The kernel mean  $m_{Q_Y}$  of  $Q_Y$  with a Gaussian kernel  $k_{R_Y}(y_1, y_2) = d_G(y_1 - y_2; 0, R_Y)$  then has the analytical solution:

$$m_{Q_Y} = \sum_{i=1}^L \xi_i d_G(\cdot; A\mu_i, R_Y + \Sigma + AW_iA^\top). \quad (15)$$

Proposition A12 describes the RKHS norm error of the Non-KSR and Mb-KSR estimators in this case.

In the experiments, we set  $m = 2$  and  $A = \Sigma = I_2$ , and drew a sample  $(X_i, Y_i)_{i=1}^{500}$  *i.i.d.* with a uniform input distribution  $P_{\mathcal{X}}$  on a square  $[-10, 10]^2$ . For the test distribution  $\Pi$ , we set  $L = 4$ ,  $\xi_i = 1/4$  ( $i = 1, 2, 3, 4$ ),  $\mu_1 = [4, 5]^\top$ ,  $\mu_2 = [-3, -5]^\top$ ,  $\mu_3 = [-6, 4]^\top$ ,  $\mu_4 = [5, -4]^\top$ ,  $W_1 = \begin{pmatrix} 2 & 0 \\ 0 & .5 \end{pmatrix}$  and  $W_i = I_2$  ( $i = 2, 3, 4$ ). The input kernel mean  $m_\Pi$  was estimated by  $\hat{m}_\Pi = \frac{1}{500} \sum_{i=1}^{500} k_{\mathcal{X}}(\cdot, \tilde{X}_i)$  with a sample  $\tilde{X}_1 \dots \tilde{X}_{500} \stackrel{i.i.d.}{\sim} \Pi$ . We set  $R_{\mathcal{X}} = 0.1I_2$ ,  $R_Y = I_2$  for the Gaussian kernels  $k_{R_{\mathcal{X}}}$  and  $k_{R_Y}$ , respectively. The results were averaged over 30 experiments.

Fig. 3 (left-upper) shows the computed averaged RKHS norm error  $\|m_{Q_Y} - \hat{m}_{Q_Y}\|_{\mathcal{H}_Y}$  vs the regularization parameter  $\epsilon$  ( $= [.1, .05, .01, .005, .001, .0005, .0001, .00005]$ ), which needs to be set for the Non-KSR. The Mb-KSR does not have a regularization parameter. 'Mb-KSR(est)' denotes the Mb-KSR where a linear Gaussian model  $(A, \Sigma)$  was learned from the training sample. This is straightforward if a probabilistic model describes the true  $P_{Y|\mathcal{X}}$  well, but it is better to learn the probabilistic model first and use it as the Mb-KSR. Fig. 3 (top right and bottom left) shows the RKHS norm errors when the degree of the model misspecification was varied for the Mb-KSR. Fig. 3 (top right) plots the errors when the Mb-KSR



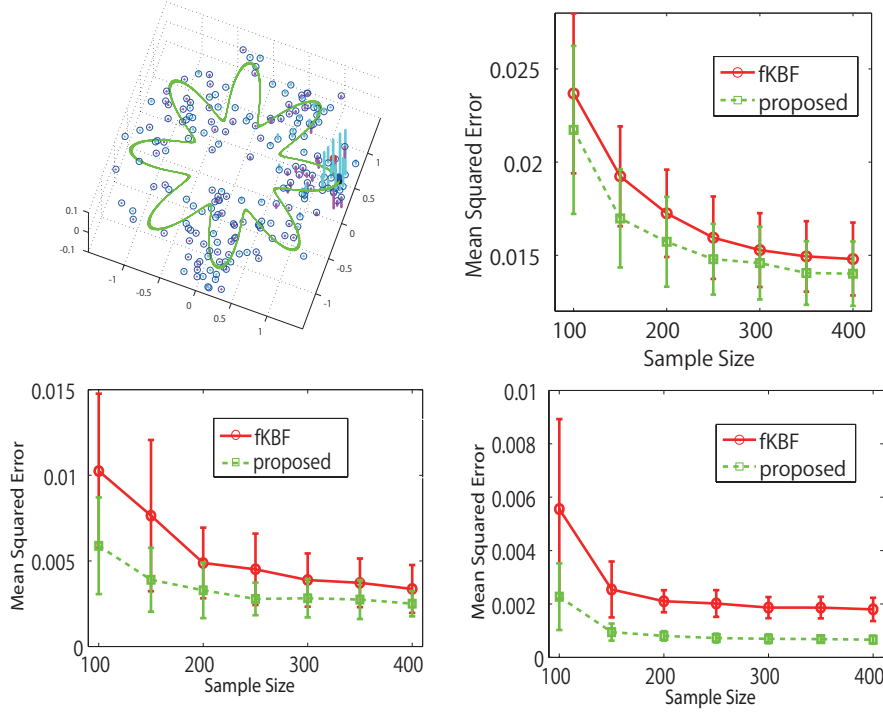
**Fig. 3** Error bars indicate standard deviations. Top left: error  $\|m_{Q_Y} - \hat{m}_{Q_Y}\|_{\mathcal{H}_Y}$  vs regularization parameter  $\epsilon$ . Top right: model misspecification, error vs scale parameter  $\sigma_1 > 0$ . Bottom left: model misspecification, error vs scale parameter  $\sigma_2 > 0$ . Bottom right: Combinations of the Non-KSR and Mb-KSR.

misspecified the coefficient  $A$  by  $\tilde{A} = \sigma_1 A$  with fixed  $\Sigma$ , where the value  $\sigma_1 > 0$  indicates the horizontal axis.  $\sigma_1 = 1$  corresponds to the exact case. Similarly, Fig. 3 (bottom left) plots the errors when the Mb-KSR misspecified the variance  $\Sigma$  by  $\tilde{\Sigma} = \sigma_2 \Sigma$  with fixed  $A$ , where the value  $\sigma_2 > 0$  indicates the horizontal axis.  $\sigma_2 = 1$  corresponds to the exact case. These figures show the sensitivity of the Mb-KSR to model misspecification. The Mb-KSR performed better than the Non-KSR with some model misspecification in this setting.

Next, we determined the inference that corresponded to eqs (11), (12). We also considered the case where the output kernel mean  $m_{Q_Z}$  had an analytical solution and where the RKHS norm error  $\|m_{Q_Z} - \hat{m}_{Q_Z}\|_{\mathcal{H}_Z}$  could be evaluated.

Let  $P_{Y|X}$  and  $P_{Z|Y}$  be the linear Gaussian noise models  $y = A_1 x + \epsilon_1$  and  $z = A_2 y + \epsilon_2$  with  $x, y, z \in \mathbb{R}^m$ , matrix  $A_1, A_2 \in \mathbb{R}^{m \times m}$  and independent Gaussian noises  $\epsilon_1 \sim N(\mathbf{0}, \Sigma_1)$ ,  $\epsilon_2 \sim N(\mathbf{0}, \Sigma_2)$ , respectively. If  $\Pi$  is a Gaussian mixture with density  $\pi(x) = \sum_{i=1}^L \xi_i d_G(x; \mu_i, W_i)$ , then the output distribution  $Q_Z$  is also a Gaussian mixture with density  $q(z) = \int p(z|y)p(y|x)\pi(x)dx dy = \sum_{i=1}^L \xi_i d_G(z; A_2 A_1 \mu_i, \Sigma_2 + A_2(\Sigma_1 + A_1 W_i A_1^\top) A_2^\top)$ . The kernel mean  $m_{Q_Z}$  of  $Q_Z$  with a Gaussian kernel  $k_{R_Z}$  has the analytical solution:

$$m_{Q_Z} = \sum_{i=1}^L \xi_i d_G(\cdot; A_2 A_1 \mu_i, R_Z + \Sigma_2 + A_2(\Sigma_1 + A_1 W_i A_1^\top) A_2^\top). \quad (16)$$



**Fig. 4** Filtering on state space models. Error bars indicate standard deviations. Experimental settings: (a) top right,  $b = 0.4$ ,  $M = 8$ ,  $\eta = 1$ ,  $\sigma_h = 0.2$ ,  $\sigma_o = 0.05$ ; (b) bottom left, Gaussian mixture noise model  $\zeta_t \sim \frac{1}{4} \sum_{i=1}^4 N(\mu_i, (0.3)^2 I_2)$  with  $\mu_1 = (0.2, 0.2)^\top$ ,  $\mu_2 = (0.2, -0.2)^\top$ ,  $\mu_3 = (-0.2, 0.2)^\top$ ,  $\mu_4 = (-0.2, -0.2)^\top$ , and  $b = 0.4$ ,  $M = 8$ ,  $\eta = 1$ ,  $\sigma_o = 0.01$ ; (c) bottom right, time-variant state space model,  $\eta = 0.1$  for training,  $\eta = 0.4$  for testing, and  $b = 0.4$ ,  $M = 8$ ,  $\sigma_h = 0.1$ ,  $\sigma_o = 0.01$ .

Proposition A13 describes the RKHS norm errors of the three estimators, i.e., (i), (ii), and (iii), in Example 32 in this setting.

In the experiments, we set  $A_1 = A_2$ ,  $\Sigma_1 = \Sigma_2$  and used the same parameter setting as that given above. Fig. 3 (bottom right) shows the errors of the three types of estimators in Proposition A13. 'N+N,' 'N+Mb,' and 'Mb+N' represent (i) Non-KSR + Non-KSR, (ii) Non-KSR + Mb-KSR, and (iii) Mb-KSR + Non-KSR in Proposition A13, respectively. N+eMb indicates the Non-KSR + Mb-KSR(est). All of the errors decreased with the regularization parameter  $\epsilon \rightarrow 0$ . We can see that estimators (ii) and (iii) were more accurate than (i) because they reflected partial knowledge ( $P_{\mathcal{Y}|\mathcal{X}}$  or  $P_{\mathcal{Z}|\mathcal{Y}}$ ) included in the linear Gaussian noise model in the Mb-KSR.

## 5.2 Filtering for state space models

We performed experiments using the proposed kernel mean filter and Algorithm 1 on a state space model (Fukumizu et al, 2013, Sect. 5.3) and compared the results



with those obtained using the existing kernel mean filter (Fukumizu et al, 2013, Sect. 4.3). The model settings were as follows.

- A hidden state was  $x_t := (u_t, v_t) \in \mathbb{R}^2$ . The transition dynamics were  $(u_{t+1}, v_{t+1})^\top = (1 + b \sin(M\theta_{t+1}))(\cos \theta_{t+1}, \sin \theta_{t+1})^\top + \varsigma_t$ ,  $\theta_{t+1} = \theta_t + \eta \pmod{2\pi}$ , where  $\varsigma_t$  was an independent Gaussian noise  $N(\mathbf{0}, \sigma_h^2 I_2)$ .
- An observation variable was  $z_t \in \mathbb{R}^2$ . A training sample  $(x_i, z_i)_{i=1}^n$  was used for learning the unknown observation process<sup>3</sup>.

We used Gaussian kernels for the state and observation domains. We learned the kernel parameters (regularization parameters  $\delta, \epsilon$  and band width parameters  $\Sigma_{\mathcal{X}} = \sigma_{\mathcal{X}}^2 I_2$ ,  $\Sigma_{\mathcal{Z}} = \sigma_{\mathcal{Z}}^2 I_2$ ) based on a twofold cross validation with a grid search. In the test phase, we estimated a single hidden state from a posterior kernel mean using eq. (14). We then evaluated the mean squared error (MSE) of the estimated state trajectory relative to the true state trajectory in a test sequence and we computed the average MSE based on 30 experiments.

Fig. 4 (top left) shows the posterior kernel mean estimator, where the green curve indicates the transition function  $f$  without the noise term and the dots indicate the training sample  $\{x_i\}_{i=1}^n$ . For each time point  $t$ , the kernel mean of the posterior distribution was estimated from the weights using the data  $\{x_i\}_{i=1}^n$ . Cyan-colored data indicate positive weights and magenta-colored data indicate negative weights. Fig. 4 (top right) shows the computed MSEs and their standard deviations as a function of the training sample size  $n$ . The parameter settings are shown in the caption. In this figure, 'proposed' indicates the results obtained with Algorithm 1 and "fKBF" indicates the results obtained using the existing full nonparametric kernel Bayes' filter (Fukumizu et al, 2013, Sect. 4.3). fKBF learns both the transition dynamics and observation process nonparametrically using transition samples  $\{(x_t, x_{t+1})\}_{t=1}^n$  and observation samples  $(x_i, z_i)_{i=1}^n$ . Since the fKBF is known to perform better than both the extended Kalman filter and unscented Kalman filter when the state space models have strong nonlinearity (Fukumizu et al, 2013, Sect. 5.3), we only report the comparison between Algorithm 1 and fKBF. We can see that the proposed Algorithm 1 may obtain a more accurate MSE than fKBF due to the incorporation of the Mb-KSR. Fig. 4 (bottom left) shows that similar results were obtained when the transition dynamics noise was extended to a Gaussian mixture. Algorithm 1 can also be applied to an additive Gaussian mixture noise model, which is obtained by a simple extension (Appendix A.1, Example A9). Another benefit of Algorithm 1 over fKBF is that Algorithm 1 can respond to time-variant (inhomogeneous) transition dynamics such as  $x_{t+1} = f_t(x_t) + \varsigma_t$ , whereas the existing fKBF cannot. Fig. 4 (bottom right) shows the simplest case where the transition dynamics of the training and test phases are different, which demonstrates fKBF cannot deal with this situation and it exhibits bias errors. The nonparametric learning of time-variant transition dynamics  $P_{\mathcal{X}'|\mathcal{X}}^{(t)}$  would require a training sample  $\{(x, x')\}_{i=1}^n$  for each fixed time  $t$ , thereby demanding a vast number of training samples.

<sup>3</sup> The actual observation process was set to  $z_t = (\text{sign}(u_t)|u_t|^{\frac{1}{2}}, \text{sign}(v_t)|v_t|^{\frac{1}{2}}) + \xi_t$ , where  $\text{sign}(\cdot)$  is the sign function and  $\xi_t \sim \text{Lap}(\mathbf{0}, \epsilon_L^{-1} I_2)$  is a Laplace noise with standard deviation  $\sigma_o = \sqrt{2}\epsilon_L^{-1}$ . We generated a training sample  $(x_i, z_i)_{i=1}^n$  accordingly. The transition dynamics and observation process were nonlinear.

### 5.3 Robot Localization

We applied the proposed Algorithm 1 to a vision-based robot localization problem. This task required the sequential estimation of the position of a mobile robot based on observations of images captured by the robot. In the state space modeling, the state domain  $\mathcal{X}$  was the robot’s position and the observation domain  $\mathcal{Z}$  comprised the images. Note that Algorithm 1 permits an arbitrary domain  $\mathcal{Z}$ , such as images, if a positive-definite kernel is defined.

In this experiment, we used the COsy Localization Database (COLD) (Pronobis and Caputo, 2009), which contains image sequences captured by mobile robots in indoor laboratory environments. Each image  $z$  was labeled according to the robot’s position  $(x, y, \theta) \in \mathbb{R}^2 \times [-\pi, \pi]$ , where  $(x, y)$  denotes the position in a building (global coordinate system) and  $\theta$  is an angle that represents the pose of the robot, and the robot’s internal odometry data  $(\bar{x}, \bar{y}, \bar{\theta})$ , i.e., the data for each time point  $t$  comprised  $D_t = (x_t, y_t, \theta_t, \bar{x}_t, \bar{y}_t, \bar{\theta}_t, z_t)$ . The process used to observe an image  $z$  given the robot’s position  $(x, y, \theta)$  is complex and it depends on the environment. Thus, we employed nonparametric learning using kernel means. However, the transition dynamics of the robot’s position  $(x, y, \theta)$  could be modeled using a probabilistic odometry motion model (Thrun et al, 2005).

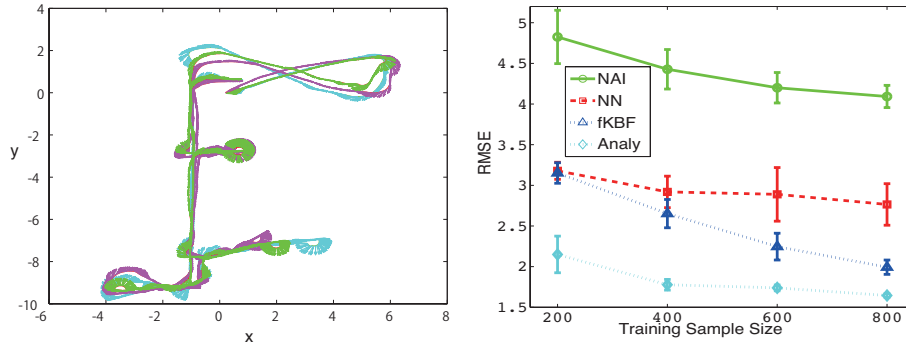
In the experiments, we used the datasets *Saarbrücken, Part A, Standard, and Cloudy* from the database, which comprised three similar trajectories on the path (Pronobis and Caputo, 2009, p. 590, Fig.1(b), blue). Fig. 5 (left) shows the trajectories with arrows. We used two of these datasets for training and the others for testing. We used the following odometry motion model to predict the next state  $(x', y', \theta')$  given the current state  $(x, y, \theta)$ , the current odometry measurement  $(\bar{x}, \bar{y}, \bar{\theta})$ , and the next odometry measurement  $(\bar{x}', \bar{y}', \bar{\theta}')$ :

$$\begin{aligned} x' &= x + \delta_{trans} \cos(\theta + \delta_{rot1}) + \xi_x, & \delta_{rot1} &= \text{atan2}(\bar{y}' - \bar{y}, \bar{x}' - \bar{x}) - \bar{\theta}, \\ y' &= y + \delta_{trans} \sin(\theta + \delta_{rot1}) + \xi_y, & \delta_{trans} &= ((\bar{x}' - \bar{x})^2 + (\bar{y}' - \bar{y})^2)^{\frac{1}{2}}, \\ \cos \theta' &= \cos(\theta + \delta_{rot1} + \delta_{rot2}) + \xi_c, & \delta_{rot2} &= \bar{\theta}' - \bar{\theta} - \delta_{rot1}, \\ \sin \theta' &= \sin(\theta + \delta_{rot1} + \delta_{rot2}) + \xi_s, \end{aligned}$$

where  $\xi_x \sim N(0, \sigma_x^2)$ ,  $\xi_y \sim N(0, \sigma_y^2)$ ,  $\xi_c \sim N(0, \sigma_c^2)$ , and  $\xi_s \sim N(0, \sigma_s^2)$  are Gaussian noise. Here,  $\text{atan2}(\cdot, \cdot)$  is an arctangent function with two arguments. We learned the variances  $\sigma_x^2$ ,  $\sigma_y^2$ ,  $\sigma_c^2$ , and  $\sigma_s^2$  using the two training datasets. We used the spatial pyramid matching kernel (Lazebnik et al, 2006) based on scale-invariant feature transform (SIFT) descriptors (Lowe, 2004) for images  $\mathcal{Z}$  and a Gaussian kernel for states  $(x, y, \cos \theta, \sin \theta)$ . The bandwidth parameters and regularization parameters were tuned based on twofold cross-validations using the training datasets. In the test phase, we estimated a filtered state as a training data point that maximized the weight of the posterior kernel mean. We then evaluated the root mean squared error (RMSE) of a test sequence and the result was averaged over 10 experiments for each training dataset size.

We compared the results obtained using the following methods.

- Naïve method (NAI): Given a test image  $z$ , the position  $(x, y, \theta)$  of a robot was estimated as a point in the training dataset that maximized the similarity in terms of the same kernel with spatial pyramid matching. Thus, this method did not consider the Markov property of the time series.



**Fig. 5** Left; three similar trajectories (colored) of a mobile robot. Each state  $(x, y, \theta)$  (arrows represent  $\theta$ ) is associated with an image. Right; RMSE vs training sample size. Error bars indicate standard deviations.

- Nearest Neighbors (NN): A  $k$  nearest neighbor approach for filtering the observation process with nonparametric learning (Vlassis et al, 2002). Given a test image  $z$ ,  $k$  nearest neighbors of the training image were explored and the likelihood function
- fKBF (Fukumizu et al, 2013): this algorithm estimated the hidden states based on nonparametric learning of both the observation process and the transition dynamics. The number of training samples used to determine the transition dynamics was fixed at 900. This comparison demonstrated the effect of incorporating the odometry motion model.

Fig. 5 (right) shows the RMSEs computed with different training sample dataset sizes. NN performed better than NAI because NAI did not consider the Markov property. The proposed method (Analy) yielded more accurate RMSEs than the fKBF by combining the odometry motion models, which was achieved by introducing the Mb-KSR.

## 6 Conclusion

In this study, we proposed a kernel Bayesian inference method that combines non-parametric estimation of conditional distributions and probabilistic modeling of conditional distributions, such as additive Gaussian noise models, in a flexible manner. This was achieved by introducing the Mb-KSR and by combining rules, i.e., Non-KSR, Mb-KSR, and KBR. We demonstrated the consistency of the Mb-KSR and showed how to combine the Non-KSR and Mb-KSR in a chain example. We also developed a filtering algorithm for state space models that combines non-parametric learning of the observation process using kernel means and additive Gaussian noise models of the transition dynamics. The idea of the Mb-KSR can be extended to  $\alpha$ -stable noise models or to elliptical distribution cases, by exploiting the concept of a conjugate pair that comprises a positive-definite kernel and a probabilistic model. In contrast to the Non-KSR, the Mb-KSR does not contain a regularization parameter, which means that tuning is not necessary. The proposed algorithm, Algorithm 1, can deal with filtering for state space models where the

transition dynamics are even time-variant, which is not possible with the current full nonparametric KBR filter. Another potential application of the Mb-KSR may be learning a probabilistic model in the setting of kernel Bayesian inference. This could be achieved in future research by learning the parameters of the transition dynamics partly by maximizing the likelihood given a sequence of observations, possibly with expectation-maximization-like algorithms. Learning during kernel Bayesian inference with probabilistic models may allow semi-parametric inference in the kernel method.

## A Appendix

### A.1 Conditional Kernel Means of Probabilistic Models

This appendix describes examples of noise models for Mb-KSR in addition to additive Gaussian noise models, which are based on the results of Nishiyama and Fukumizu (2014).

Nishiyama and Fukumizu (2014) described a systematic approach for determining the kernel means of probabilistic models in terms of infinitely divisible distributions, where the conjugate pair of a probabilistic model  $p$  and a positive-definite kernel  $k$  was introduced to ensure that its kernel mean  $m_p$  had a simple form. A probabilistic model  $p$  and positive-definite kernel  $k$  are called *conjugate* if  $p$  and  $m_p$  has the same density function form. If  $k$  is shift invariant, the kernel mean map  $p \mapsto m_p$  implies a convolution  $p \mapsto \psi * p$  with the positive-definite function  $\psi$ . Thus, the conjugate property indicates the *reproducing property of probability distributions* (i.e., a family of distributions that are closed under convolution) by the inclusion of a positive-definite function  $\psi$ .

For example, the family of Gaussian distributions is closed under convolution and it includes a Gaussian density function  $\psi$  that is positive-definite. The kernel mean  $\psi * p$  of a Gaussian density  $p$  with a Gaussian kernel  $\psi(x - y)$  is then a Gaussian. More generally, for each  $\alpha \in (0, 2]$ , the family of  $\alpha$ -stable distributions is closed under convolution and it includes an  $\alpha$ -stable density function  $\psi$  that is positive-definite<sup>4</sup> (Nishiyama and Fukumizu, 2014, Sect. 4, p. 16).  $\alpha$ -stable distributions have heavy tails for  $\alpha \in (0, 2)$  and light tails for  $\alpha = 2$  (Gaussian cases), and  $\alpha$  tunes the degree of the tail property of the distribution and the positive-definite kernel. By contrast, the family of Gamma distributions is closed under convolution but it does not include a Gamma density  $\psi$  that is positive-definite, and this is not a conjugate case. The family of Laplace distributions is not closed under convolution and this is not a conjugate case. However, note that the conditional kernel mean of a Laplace distribution with a Laplace kernel is given by an explicit form (Example A6) and it can be used for the Mb-KSR.

The results obtained using the kernel means of probabilistic models (Nishiyama and Fukumizu, 2014) immediately indicate the conditional kernel mean cases for additive probabilistic noise models.

The  $\alpha$ -stable case on  $\mathbb{R}$  is as follows.

Let  $d_{S_\alpha}(x|\sigma, \beta, \mu)$ ,  $x \in \mathbb{R}$  denote the  $\alpha$ -stable density on  $\mathbb{R}$ , where  $\alpha \in (0, 2]$  is a *characteristic index*,  $\sigma > 0$  is a scale parameter,  $\beta \in \mathbb{R}$  is a skewness parameter, and  $\mu \in \mathbb{R}$  is a location parameter. We denote an  $\alpha$ -stable random variable  $X$  as  $X \sim S_\alpha(\sigma, \beta, \mu)$ . Let  $\mathcal{H}_{\alpha, \sigma}$  be the RKHS generated by an  $\alpha$ -stable kernel  $k_{\alpha, \sigma}(x, y) = d_{S_\alpha}(x - y|\sigma, 0, 0)$ ,  $x, y \in \mathbb{R}$ .

**Example A1 (Additive  $\alpha$ -stable noise model on  $\mathbb{R}$ )** Let  $\mathcal{Y} = \mathbb{R}$ . Assume that  $\{P_{\mathcal{Y}|x}|x \in \mathcal{X}\}$  is an additive  $\alpha$ -stable noise model  $y = f(x) + \epsilon$  with a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  and an  $\alpha$ -stable noise  $\epsilon \sim S_\alpha(\sigma, \beta, 0)$ . The conditional kernel mean  $m_{\mathcal{Y}|x}$  of the additive  $\alpha$ -stable noise model in the  $\alpha$ -stable RKHS  $\mathcal{H}_{\alpha, \sigma_0}$  is an  $\alpha$ -stable density, which is function given by

$$m_{\mathcal{Y}|x} = d_{S_\alpha}(\cdot | (\sigma_0^\alpha + \sigma^\alpha)^{\frac{1}{\alpha}}, \frac{\sigma^\alpha \beta}{\sigma_0^\alpha + \sigma^\alpha}, f(x)) \in \mathcal{H}_{\alpha, \sigma_0}. \quad (17)$$

The generalization to multi-dimensional  $\alpha$ -stable cases on  $\mathbb{R}^m$  is similar.

---

<sup>4</sup> The Gaussian case corresponds to  $\alpha = 2$ .

Let  $d_{S_\alpha}(x|\Gamma, \mu)$ ,  $x \in \mathbb{R}^m$  denote the multi-dimensional  $\alpha$ -stable density on  $\mathbb{R}^m$ , where  $\alpha \in (0, 2]$  is a characteristic index,  $\mu \in \mathbb{R}^m$  is a location parameter, and  $\Gamma$  is a *spectral measure* on the unit sphere  $S_{m-1} := \{s \in \mathbb{R}^m : \|s\| = 1\}$ . We denote an  $\alpha$ -stable random vector  $X$  on  $\mathbb{R}^m$  as  $X \sim S_\alpha(\Gamma, \mu)$ . Let  $\mathcal{H}_{\alpha, \Gamma_s}$  be the RKHS generated by an  $\alpha$ -stable kernel  $k_{\alpha, \Gamma_s}(x, y) = d_{S_\alpha}(x - y|\Gamma_s, 0)$ ,  $x, y \in \mathbb{R}^m$ , where  $\Gamma_s$  is a symmetric spectral measure (i.e.,  $\Gamma_s(A) = \Gamma_s(-A)$  for any  $A \in \mathcal{B}(S_{m-1})$ ).

**Example A2 (Additive  $\alpha$ -stable noise model on  $\mathbb{R}^m$ )** Let  $\mathcal{Y} = \mathbb{R}^m$ . Assume that  $\{P_{\mathcal{Y}|x}|x \in \mathcal{X}\}$  is an additive  $\alpha$ -stable noise model  $y = f(x) + \epsilon$  with a function  $f : \mathcal{X} \rightarrow \mathbb{R}^m$  and an  $\alpha$ -stable random vector noise  $\epsilon \sim S_\alpha(\Gamma, 0)$ . The conditional kernel mean  $m_{\mathcal{Y}|x}$  of the additive  $\alpha$ -stable noise model in the  $\alpha$ -stable RKHS  $\mathcal{H}_{\alpha, \Gamma_s}$  is an  $\alpha$ -stable density function given by

$$m_{\mathcal{Y}|x} = d_{S_\alpha}(\cdot|\Gamma_s + \Gamma, f(x)) \in \mathcal{H}_{\alpha, \Gamma_s}. \quad (18)$$

A well-known special case of the multi-dimensional  $\alpha$ -stable distributions on  $\mathbb{R}^m$  comprises *sub-Gaussian*  $\alpha$ -stable distributions on  $\mathbb{R}^m$ .

Let  $d_{SG_\alpha}(x|R, \mu)$ ,  $x \in \mathbb{R}^m$  denote the sub-Gaussian  $\alpha$ -stable density on  $\mathbb{R}^m$ , where  $\alpha \in (0, 2)$  is a characteristic index,  $\mu \in \mathbb{R}^m$  is a location parameter, and  $R \in \mathbb{R}^{m \times m}$  is a dispersion matrix. We denote a sub-Gaussian  $\alpha$ -stable random vector  $X$  on  $\mathbb{R}^m$  as  $X \sim SG_\alpha(R, \mu)$ . Let  $\mathcal{H}_{\alpha, R}$  be the RKHS generated by an  $\alpha$ -stable kernel  $k_{\alpha, R}(x, y) = d_{SG_\alpha}(x - y|R, 0)$ ,  $x, y \in \mathbb{R}^m$ .  $\alpha = 2$  corresponds to multi-dimensional Gaussian cases.

**Example A3 (Additive sub-Gaussian  $\alpha$ -stable noise model on  $\mathbb{R}^m$ )** Let  $\mathcal{Y} = \mathbb{R}^m$ . Assume that  $\{P_{\mathcal{Y}|x}|x \in \mathcal{X}\}$  is an additive sub-Gaussian  $\alpha$ -stable noise model  $y = f(x) + \epsilon$  with a function  $f : \mathcal{X} \rightarrow \mathbb{R}^m$  and  $\epsilon \sim SG_\alpha(0, R)$ . If  $R_0 = cR$  with  $c > 0$ , the conditional kernel mean  $m_{\mathcal{Y}|x}$  of the additive sub-Gaussian  $\alpha$ -stable noise model in the sub-Gaussian  $\alpha$ -stable RKHS  $\mathcal{H}_{\alpha, R_0}$  is a sub-Gaussian  $\alpha$ -stable density function, which is given by

$$m_{\mathcal{Y}|x} = d_{SG_\alpha}(\cdot|f(x), (c^{\frac{\alpha}{2}} + 1)^{\frac{2}{\alpha}} R_0) \in \mathcal{H}_{\alpha, R_0}. \quad (19)$$

Gaussian distributions and sub-Gaussian  $\alpha$ -stable distributions are examples of *elliptical distributions*. More conceptually, the results given above can be extended to general elliptical distributions. Appendix B reviews elliptical distributions.

We denote  $\mathbb{R}_{\geq 0} = [0, \infty)$ . Let  $d_{EC}(x|\mu, \Sigma, \phi)$ ,  $x \in \mathbb{R}^m$  be the elliptical density on  $\mathbb{R}^m$ , where  $\mu \in \mathbb{R}^m$  is a location parameter,  $\Sigma \in \mathbb{R}^{m \times m}$  is a dispersion matrix, and  $\phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{C}$  is a *characteristic generator*. We denote an elliptical random vector  $X$  on  $\mathbb{R}^m$  as  $X \sim EC(\mu, \Sigma, \phi)$ . Let  $\mathcal{H}_{\Sigma, \phi_0}$  be the RKHS generated by an elliptical density kernel  $k_{\Sigma, \phi_0}(x, y) = d_{EC}(x - y|0, \Sigma, \phi_0)$ ,  $x, y \in \mathbb{R}^m$ , where  $\phi_0 : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  is a nonnegative characteristic generator. The Gaussians distributions and sub-Gaussian  $\alpha$ -stable distributions correspond to  $\phi(r) = e^{-\frac{1}{2}r}$  and  $\phi(r) = e^{-\frac{1}{2}r^{\frac{\alpha}{2}}}$ , respectively.

**Example A4 (Additive elliptical noise model on  $\mathbb{R}^m$ )** Let  $\mathcal{Y} = \mathbb{R}^m$ . Assume that  $\{P_{\mathcal{Y}|x}|x \in \mathcal{X}\}$  is an additive elliptical noise model  $y = f(x) + \epsilon$  with a function  $f : \mathcal{X} \rightarrow \mathbb{R}^m$  and  $\epsilon \sim EC(0, \Sigma, \phi)$ . The conditional kernel mean  $m_{\mathcal{Y}|x}$  of the additive elliptical noise model in the RKHS  $\mathcal{H}_{\Sigma, \phi_0}$  is an elliptical density function given by

$$m_{\mathcal{Y}|x} = d_{EC}(\cdot; f(x), \Sigma, \tilde{\phi}) \in \mathcal{H}_{\Sigma, \phi_0}, \quad (20)$$

where  $\tilde{\phi}(r) = \phi(r)\phi_0(r)$ .

*Proof* For each  $x \in \mathcal{X}$ ,  $m_{\mathcal{Y}|x}$  is given by a convolution  $m_{\mathcal{Y}|x} = d_{EC}(\cdot|0, \Sigma, \phi_0) * d_{EC}(\cdot|f(x), \Sigma, \phi)$ . In general, we use the following lemma, which completes the proof.

**Lemma A5** (McNeil et al, 2005, p. 95) If  $X_1$  and  $X_2$  are independent elliptical random vectors  $X_1 \sim EC(\mu_1, \Sigma, \phi_1)$  and  $X_2 \sim EC(\mu_2, \Sigma, \phi_2)$  on  $\mathbb{R}^m$ , then  $X + Y$  is elliptical:

$$X + Y \sim EC(\mu_1 + \mu_2, \Sigma, \tilde{\phi}), \quad (21)$$

where  $\tilde{\phi}(r) = \phi_1(r)\phi_2(r)$ .

Normal variance mixture distributions are examples of elliptical distributions other than the Gaussian and sub-Gaussian cases. The density is given by a scale mixture of Gaussian densities  $\int d_G(x|\mu, w\Sigma)dH(w)$  with a mixing distribution  $H$  on  $\mathbb{R}_{>0}$  (McNeil et al, 2005, Definition 3.4, p. 73). The characteristic generator is given by  $\phi(r) = \hat{H}(\frac{1}{2}r)$ , where  $\hat{H}(\theta)$  is the Laplace-Stieltjes transform  $\hat{H}(\theta) := \int_0^\infty e^{-\theta w} dH(w)$  of the mixing distribution  $H$  (McNeil et al, 2005, Example 3.21, p. 90). If  $H$  is a generalized inverse Gaussian distribution and the normal variance mixture distribution is an elliptical *generalized hyperbolic (GH) distribution*  $GH_d(\lambda, \chi, \psi, \mu, \Sigma, \gamma = 0)$  (McNeil et al, 2005, Example 3.8, p. 75). The elliptical normal inverse Gaussian (NIG)  $GH_d(\lambda = -\frac{1}{2}, \chi, \psi, \mu, \Sigma, \gamma = 0)$  and elliptical variance gamma (VG) distributions  $GH_d(\lambda > 0, \chi = 0, \psi, \mu, \Sigma, \gamma = 0)$  are subclasses of the elliptical GH distributions (McNeil et al, 2005, Sect. 3.2.3, p. 78). The elliptical NIG and VG have the same conjugate property as that given above (Nishiyama and Fukumizu, 2014).

The Laplace distribution case is described as follows.

Let  $d_L(x|\mu, \lambda) := \frac{\lambda}{2} \exp(-\lambda|x - \mu|)$ ,  $x \in \mathbb{R}$  denote the Laplace density on  $\mathbb{R}$ . We denote a Laplace random variable  $X$  on  $\mathbb{R}$  as  $X \sim \text{Lap}(\mu, \lambda)$ . Let  $\mathcal{H}_\lambda^{(L)}$  be the Laplace RKHS generated by a Laplace kernel  $k_\lambda(x_1, x_2) = d_L(x_1 - x_2|0, \lambda)$ ,  $x_1, x_2 \in \mathbb{R}$ .

**Example A6 (Additive Laplace noise model on  $\mathbb{R}$ )** Let  $\mathcal{Y} = \mathbb{R}$ . Assume that  $\{P_{\mathcal{Y}|x}|x \in \mathcal{X}\}$  is an additive Laplace noise model  $y = f(x) + \epsilon$  with a function  $f: \mathcal{X} \rightarrow \mathbb{R}$  and a Laplace noise  $\epsilon \sim \text{Lap}(0, \lambda)$ . The conditional kernel mean  $m_{\mathcal{Y}|x}$  of the additive Laplace noise model in a Laplace RKHS  $\mathcal{H}_{\lambda_0}^{(L)}$  for each  $y \in \mathbb{R}$  is given by

$$m_{\mathcal{Y}|x}(y) = \begin{cases} \frac{\lambda_0 \lambda (\lambda e^{-\lambda_0|y-f(x)|} - \lambda_0 e^{-\lambda|y-f(x)|})}{2(\lambda^2 - \lambda_0^2)} & (\lambda \neq \lambda_0) \\ \frac{\lambda_0(1 + \lambda_0|y-f(x)|)e^{-\lambda_0|y-f(x)|}}{4} & (\lambda = \lambda_0). \end{cases} \quad (22)$$

*Proof* It suffices to show the following and we then derive Proposition A7.

**Proposition A7 (Laplace distribution + Laplace kernel)** Let  $\mathcal{Y} = \mathbb{R}$ . The kernel mean  $m_{\mathcal{Y}}$  of a Laplace density  $d_L(y|\mu, \lambda)$ ,  $y \in \mathbb{R}$  in a Laplace RKHS  $\mathcal{H}_{\lambda_0}^{(L)}$  for each  $y \in \mathbb{R}$  is given by,

$$m_{\mathcal{Y}}(y) = \begin{cases} \frac{\lambda_0 \lambda (\lambda e^{-\lambda_0|y-\mu|} - \lambda_0 e^{-\lambda|y-\mu|})}{2(\lambda^2 - \lambda_0^2)} & (\lambda \neq \lambda_0) \\ \frac{\lambda_0(1 + \lambda_0|y-\mu|)e^{-\lambda_0|y-\mu|}}{4} & (\lambda = \lambda_0). \end{cases} \quad (23)$$

*Proof* By definition,  $m_{\mathcal{Y}}(y)$ ,  $y \in \mathbb{R}$  is given by

$$\begin{aligned} m_{\mathcal{Y}}(y) &= \int_{-\infty}^{\infty} \frac{\lambda}{2} e^{-\lambda|\tilde{y}-\mu|} d_L(y - \tilde{y}|\mu, \lambda) d\tilde{y} \\ &= \int_{-\infty}^{\infty} \frac{\lambda e^{-\lambda(\tilde{y}-\mu)} \mathbf{1}_{[\mu, \infty]} + \lambda e^{-\lambda(\mu-\tilde{y})} \mathbf{1}_{[-\infty, \mu]}}{2} d_L(y - \tilde{y}|\mu, \lambda) d\tilde{y} \\ &= \frac{1}{2} m_{\lambda, \mu}^{(e)}(y) + \frac{1}{2} m_{\lambda, -\mu}^{(e)}(-y), \end{aligned} \quad (24)$$

where  $m_{\lambda, \mu}^{(e)}$  is the kernel mean of an exponential distribution, as follows.

**Lemma A8 (Exponential distribution + Laplace kernel)** Let  $\mathcal{Y} = \mathbb{R}$ . The kernel mean  $m_{\lambda, \mu}^{(e)}$  of an exponential density  $d_e(y|\mu, \lambda) := \lambda \exp(-\lambda(y-\mu)) \mathbf{1}_{[\mu, \infty]}$ ,  $y \in \mathbb{R}$  in a Laplace RKHS  $\mathcal{H}_{\lambda_0}^{(L)}$  for  $y \in \mathbb{R}$  is given by,

$$m_{\lambda, \mu}^{(e)}(y) = \begin{cases} \frac{\lambda_0 \lambda}{2(\lambda_0 + \lambda)} e^{-\lambda_0(\mu - y)} & (y \leq \mu) \\ \frac{\lambda_0 \lambda}{2(\lambda_0 + \lambda)} e^{-\lambda(y - \mu)} + \frac{\lambda_0 \lambda}{2(\lambda_0 - \lambda)} (e^{-\lambda(y - \mu)} - e^{-\lambda_0(y - \mu)}) & (y \geq \mu, \lambda \neq \lambda_0) \\ \frac{\lambda^2}{2} e^{-\lambda(y - \mu)}(y - \mu) + \frac{\lambda}{4} e^{-\lambda(y - \mu)} & (y \geq \mu, \lambda = \lambda_0) \end{cases}$$

The proof of Lemma A8 can be obtained by direct computation and it is omitted. Eq. (24) and Lemma A8 complete the derivation.

The conditional kernel mean of an additive Laplace noise model in a Laplace RKHS has an explicit form, as given above, but it is not exactly a Laplace density function. This is not a conjugate case. However, note that the Mb-KSR with the Laplace noise model is computable using the explicit expression (23).

The Mb-KSR may also be employed with a mixture of probabilistic models for the examples above.

**Example A9 (mixtures)** Let  $\{P_{\mathcal{Y}|x}^{(k)}\}_{k=1}^N$  be  $N$  conditional distributions and  $\{m_{\mathcal{Y}|x}^{(k)}\}_{k=1}^N$  are their conditional kernel means  $\{m_{\mathcal{Y}|x}^{(k)}\}_{k=1}^N$  in a common RKHS  $\mathcal{H}_{\mathcal{Y}}$ , respectively. The conditional kernel mean of mixture  $P_{\mathcal{Y}|x} = \sum_{k=1}^N \omega_k P_{\mathcal{Y}|x}^{(k)}$  in  $\mathcal{H}_{\mathcal{Y}}$  is a mixture of the kernel means  $m_{\mathcal{Y}|x} = \sum_{k=1}^N \omega_k m_{\mathcal{Y}|x}^{(k)}$  in  $\mathcal{H}_{\mathcal{Y}}$ . If  $\{m_{\mathcal{Y}|x}^{(k)}\}_{k=1}^N$  have expressions, as in the examples above, then  $m_{\mathcal{Y}|x}$  is a mixture of them.

## A.2 Consistency

The consistency of the Mb-KSR is demonstrated as follows. The rate of the output kernel mean is the same as that of the input kernel mean  $\hat{m}_{\Pi}$ . Note that Fukumizu et al (2013, Theorem 11, p.3776) provides the consistency rate of the Non-KSR as an upper bound. The rate is always lower than that of the input kernel mean.

**Proposition A10** Suppose that a kernel mean estimator  $\hat{m}_{\Pi} := \sum_{i=1}^l w_i k_{\mathcal{X}}(\cdot, \tilde{X}_i)$  satisfies  $\|\hat{m}_{\Pi} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}} = O_p(l^{-\alpha})$  as  $l \rightarrow \infty$  for some  $0 < \alpha \leq \frac{1}{2}$ . Assume that  $\theta(x, \tilde{x}) := \int k_{\mathcal{Y}}(y, \tilde{y}) dP_{\mathcal{Y}|x}(y) dP_{\mathcal{Y}|\tilde{x}}(\tilde{y})$  is included in  $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}$  as a function of  $(x, \tilde{x})$ . Then,  $\hat{m}_{Q_{\mathcal{Y}}} := \sum_{i=1}^l w_i m_{\mathcal{Y}|\tilde{X}_i}$  is a consistent estimator with the same rate  $\|m_{Q_{\mathcal{Y}}} - \hat{m}_{Q_{\mathcal{Y}}}\|_{\mathcal{H}_{\mathcal{Y}}} = O_p(l^{-\alpha})$ .

*Proof* We have the following derivation.

$$\begin{aligned}
& \left\| m_{Q_{\mathcal{Y}}} - \sum_{i=1}^l w_i m_{\mathcal{Y}|\tilde{X}_i} \right\|_{\mathcal{H}_{\mathcal{Y}}}^2 \\
&= \sum_{i,j=1}^l w_i w_j \langle m_{\mathcal{Y}|\tilde{X}_i}, m_{\mathcal{Y}|\tilde{X}_j} \rangle_{\mathcal{H}_{\mathcal{Y}}} - 2 \sum_{i=1}^l w_i \langle m_{\mathcal{Y}|\tilde{X}_i}, m_{Q_{\mathcal{Y}}} \rangle_{\mathcal{H}_{\mathcal{Y}}} + \|m_{Q_{\mathcal{Y}}}\|_{\mathcal{H}_{\mathcal{Y}}}^2 \\
&= \sum_{i,j=1}^l w_i w_j \int k_{\mathcal{Y}}(y, \tilde{y}) dP_{\mathcal{Y}|\tilde{X}_i}(y) dP_{\mathcal{Y}|\tilde{X}_j}(\tilde{y}) \\
&\quad - 2 \sum_{i=1}^l w_i \int k_{\mathcal{Y}}(y, \tilde{y}) dP_{\mathcal{Y}|\tilde{X}_i}(y) dP_{\mathcal{Y}|x}(\tilde{y}) d\Pi(x) \\
&\quad + \int k_{\mathcal{Y}}(y, \tilde{y}) dP_{\mathcal{Y}|x}(y) dP_{\mathcal{Y}|\tilde{x}}(\tilde{y}) d\Pi(x) d\Pi(\tilde{x}) \\
&= \sum_{i,j=1}^l w_i w_j \theta(\tilde{X}_i, \tilde{X}_j) - 2 \sum_{i=1}^l w_i \int \theta(\tilde{X}_i, x) d\Pi(x) + \int \theta(x, \tilde{x}) d\Pi(x) d\Pi(\tilde{x}) \\
&= \langle (\hat{m}_{\Pi} - m_{\Pi}) \otimes (\hat{m}_{\Pi} - m_{\Pi}), \theta \rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}} \\
&\leq \|(\hat{m}_{\Pi} - m_{\Pi}) \otimes (\hat{m}_{\Pi} - m_{\Pi})\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}} \|\theta\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}} \\
&= \|\hat{m}_{\Pi} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}}^2 \|\theta\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}}
\end{aligned}$$

The fourth equality employs the assumption  $\theta(\cdot, \cdot) \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}$ . The assumption  $\|\hat{m}_{\Pi} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}} = O_p(l^{-\alpha})$  and  $\|\theta\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}} < \infty$  proves the statement.

The following are the consistency rates of the two types of estimators, i.e., (ii) and (iii), given in Example 32.

**Example A11** Suppose that a kernel mean estimator  $\hat{m}_\Pi^{(l)} := \sum_{i=1}^l \gamma_i k_\mathcal{X}(\cdot, \tilde{X}_i)$  satisfies  $\|\hat{m}_\Pi^{(l)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(l^{-\alpha})$  as  $l \rightarrow \infty$  for some  $0 < \alpha \leq \frac{1}{2}$ . Suppose that the Non-KSR has the consistency rate  $\|m_{Q_Y} - \hat{\mathcal{U}}_{Y|\mathcal{X}}^{(n)} \hat{m}_\Pi^{(l)}\|_{\mathcal{H}_Y} = O_p(n^{-c_\alpha})$  with some  $0 < c_\alpha \leq \frac{1}{2}$  as  $n, l \rightarrow \infty$ . The consistency rates of the two estimators (ii) and (iii) in Example 32 are given as follows:

- (ii)  $\|m_{Q_Z} - \hat{m}_{Q_Z}\|_{\mathcal{H}_Z} = \|m_{Q_Z} - \bar{\mathcal{U}}_{Z|Y} \hat{\mathcal{U}}_{Y|\mathcal{X}}^{(n)} \hat{m}_\Pi^{(l)}\|_{\mathcal{H}_Z} = O(n^{-c_\alpha}),$
- (iii)  $\|m_{Q_Z} - \hat{m}_{Q_Z}\|_{\mathcal{H}_Z} = \|m_{Q_Z} - \hat{\mathcal{U}}_{Z|Y}^{(n)} \hat{\mathcal{U}}_{Y|\mathcal{X}} \hat{m}_\Pi^{(l)}\|_{\mathcal{H}_Z} = O(n^{-c_\alpha}).$

An upper bound of the consistency rate of the Non-KSR is shown in Fukumizu et al (2013, Theorem 11, p.3776). If a function  $\pi/p_\mathcal{X}$  is included in the range of the  $\beta$ -th power of covariance operator  $C_{XX}$  for some  $\beta \geq 0$  (i.e.,  $\pi/p_\mathcal{X} \in \mathcal{R}(C_{XX}^\beta)$ ),  $n = l$ , and  $\epsilon_n := n^{-\max\{\frac{2}{3}\alpha, \frac{\alpha}{1+\beta}\}}$ , then  $c_\alpha$  has an upper bound  $\tilde{c}_\alpha := \min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}$ . Thus, the two estimators above have an upper bound  $O(n^{-\tilde{c}_\alpha})$ .

### A.3 RKHS norm errors in the setting described in Sect. 5.1

**Proposition A12** (i) Let  $\hat{m}_{Q_Y}$  denote the Non-KSR estimator in the setting given in Sect. 5.1. Then,

$$\|m_{Q_Y} - \hat{m}_{Q_Y}\|_{\mathcal{H}_Y}^2 = \xi^\top E\xi - 2w^\top \mathbf{m}_{Q_Y} + w^\top G_Y w, \quad (25)$$

where  $E_{ij} := d_G(0; A(\mu_i - \mu_j), R_Y + 2\Sigma + A(W_i + W_j)A^\top)$ ,  $\mathbf{m}_{Q_Y} := (m_{Q_Y}(Y_1), \dots, m_{Q_Y}(Y_n))^\top$  and the Gram matrix  $G_Y = (k_{R_Y}(Y_i, Y_j))$ .

(ii) Let  $\hat{m}_{Q_Y}$  denote the Mb-KSR estimator in the setting given in Sect. 5.1. Suppose that the Mb-KSR misspecifies the model  $(A, \Sigma)$  as  $(\tilde{A}, \tilde{\Sigma})$ . Then,

$$\|m_{Q_Y} - \hat{m}_{Q_Y}\|_{\mathcal{H}_Y}^2 = \xi^\top E\xi - 2\gamma^\top F\xi + \gamma^\top H\gamma, \quad (26)$$

where  $F_{ij} := d_G(0; \tilde{A}\tilde{X}_i - A\mu_j, R_Y + \tilde{\Sigma} + \Sigma + AW_jA^\top)$  and  $H_{ij} := d_G(0; \tilde{A}(\tilde{X}_i - \tilde{X}_j), R_Y + 2\tilde{\Sigma})$ .

*Proof* We derive (i) and (ii).

(i) First, the norm  $\|m_{Q_Y}\|_{\mathcal{H}_Y}^2$  can be computed using eq. (15) as

$$\begin{aligned} \|m_{Q_Y}\|_{\mathcal{H}_Y}^2 &= \sum_{i,j=1}^L \xi_i \xi_j \left\langle d_G(\cdot; A\mu_i, R_Y + \Sigma + AW_iA^\top), d_G(\cdot; A\mu_j, R_Y + \Sigma + AW_jA^\top) \right\rangle_{\mathcal{H}_Y} \\ &= \sum_{i,j=1}^L \xi_i \xi_j d_G(0; A(\mu_i - \mu_j), R_Y + 2\Sigma + A(W_i + W_j)A^\top) =: \xi^\top E\xi. \end{aligned} \quad (27)$$

For the second equality, we use the explicit expression of the inner product of two Gaussian kernel means (Nishiyama and Fukumizu, 2014, Proposition 4.5). Using the Non-KSR estimator  $\hat{m}_{Q_Y} = \sum_{i=1}^n w_i k(\cdot, Y_i)$ , we then obtain,

$$\begin{aligned} \|m_{Q_Y} - \hat{m}_{Q_Y}\|_{\mathcal{H}_Y}^2 &= \|m_{Q_Y}\|_{\mathcal{H}_Y}^2 - 2\langle \hat{m}_{Q_Y}, m_{Q_Y} \rangle_{\mathcal{H}_Y} + \|\hat{m}_{Q_Y}\|_{\mathcal{H}_Y}^2 \\ &= \xi^\top E\xi - 2 \left\langle \sum_{i=1}^n w_i k(\cdot, Y_i), m_{Q_Y} \right\rangle_{\mathcal{H}_Y} + \left\| \sum_{i=1}^n w_i k(\cdot, Y_i) \right\|_{\mathcal{H}_Y}^2 \\ &= \xi^\top E\xi - 2 \sum_{i=1}^n w_i m_{Q_Y}(Y_i) + w^\top G w. \end{aligned}$$



(ii) From eq. (27) and the Mb-KSR estimator  $\hat{m}_{Q_Y} = \sum_{i=1}^l \gamma_i m_{Y|\tilde{X}_i}$ ,

$$\begin{aligned}
\|m_{Q_Y} - \hat{m}_{Q_Y}\|_{\mathcal{H}_Y}^2 &= \|m_{Q_Y}\|_{\mathcal{H}_Y}^2 - 2\langle \hat{m}_{Q_Y}, m_{Q_Y} \rangle_{\mathcal{H}_Y} + \|\hat{m}_{Q_Y}\|_{\mathcal{H}_Y}^2 \\
&= \xi^\top E \xi - 2 \sum_{i=1}^l \gamma_i \langle m_{Y|\tilde{X}_i}, m_{Q_Y} \rangle_{\mathcal{H}_Y} + \sum_{i,j=1}^l \gamma_i \gamma_j \langle m_{Y|\tilde{X}_i}, m_{Y|\tilde{X}_j} \rangle_{\mathcal{H}_Y} \\
&= \xi^\top E \xi - 2 \sum_{i=1}^l \sum_{j=1}^L \gamma_i \xi_j \langle d_G(\cdot; \tilde{A}\tilde{X}_i, \tilde{\Sigma} + R_Y), d_G(\cdot; A\mu_j, R_Y + \Sigma + AW_j A^\top) \rangle_{\mathcal{H}_Y} \\
&\quad + \sum_{i,j=1}^l \gamma_i \gamma_j \langle d_G(\cdot; \tilde{A}\tilde{X}_i, \tilde{\Sigma} + R_Y), d_G(\cdot; \tilde{A}\tilde{X}_j, \tilde{\Sigma} + R_Y) \rangle_{\mathcal{H}_Y} \\
&= \xi^\top E \xi - 2 \sum_{i=1}^l \sum_{j=1}^L \gamma_i \xi_j d_G(0; \tilde{A}\tilde{X}_i - A\mu_j, R_Y + \tilde{\Sigma} + \Sigma + AW_j A^\top) \\
&\quad + \sum_{i,j=1}^l \gamma_i \gamma_j d_G(0; \tilde{A}(\tilde{X}_i - \tilde{X}_j), R_Y + 2\tilde{\Sigma}) \\
&= \xi^\top E \xi - 2\gamma^\top F \xi + \gamma^\top H \gamma,
\end{aligned}$$

which completes the derivation. For the fourth equality, we use the explicit expression of the inner product of two Gaussian kernel means (Nishiyama and Fukumizu, 2014, Proposition 4.5).

**Proposition A13** (i) (Non-KSR + Non-KSR) Let  $\hat{m}_{Q_Z}$  be an estimator  $\hat{m}_{Q_Z} = \hat{\mathcal{U}}_{Z|Y} \hat{\mathcal{U}}_{Y|X} \hat{m}_\Pi$ , where both  $\hat{\mathcal{U}}_{Z|Y}$  and  $\hat{\mathcal{U}}_{Y|X}$  are Non-KSR operations in the setting given in Sect. 5.1. Then,

$$\|m_{Q_Z} - \hat{m}_{Q_Z}\|_{\mathcal{H}_Z}^2 = \xi^\top E \xi - 2w^\top \mathbf{m}_{Q_Z} + w^\top G_Z w, \quad (28)$$

where  $w = (G_X + n\epsilon I_n)^{-1} G_{X\tilde{X}} \gamma$ ,  $E_{ij} := d_G(0; A_2 A_1 (\mu_i - \mu_j), R_Z + 2\Sigma_2 + A_2(2\Sigma_1 + A_1(W_i + W_j)A_1^\top)A_2^\top)$ ,  $\mathbf{m}_{Q_Z} := (m_{Q_Z}(Z_1), \dots, m_{Q_Z}(Z_n))^\top$  and Gram matrix  $G_Z = (k_{R_Z}(Z_i, Z_j))$ .

(ii) (Mb-KSR + Non-KSR) Let  $\hat{m}_{Q_Z}$  be an estimator  $\hat{m}_{Q_Z} = \hat{\mathcal{U}}_{Z|Y} \hat{\mathcal{U}}_{Y|X} \hat{m}_\Pi$ , where  $\hat{\mathcal{U}}_{Z|Y}$  is the Mb-KSR operation and  $\hat{\mathcal{U}}_{Y|X}$  is the Non-KSR operation in the setting given in Sect. 5.1. Then,

$$\|m_{Q_Z} - \hat{m}_{Q_Z}\|_{\mathcal{H}_Z}^2 = \xi^\top E \xi - 2w^\top F \xi + w^\top H w \quad (29)$$

where  $w = (G_X + n\epsilon I_n)^{-1} G_{X\tilde{X}} \gamma$ ,  $F_{ij} = d_G(0; A_2 Y_i - A_2 A_1 \mu_j, R_Z + 2\Sigma_2 + A_2(\Sigma_1 + A_1 W_j A_1^\top) A_2^\top)$  and  $H_{ij} = d_G(0; A_2(Y_i - Y_j), R_Z + 2\Sigma_2)$ .

(iii) (Non-KSR + Mb-KSR) Let  $\hat{m}_{Q_Z}$  be an estimator  $\hat{m}_{Q_Z} = \hat{\mathcal{U}}_{Z|Y} \hat{\mathcal{U}}_{Y|X} \hat{m}_\Pi$ , where  $\hat{\mathcal{U}}_{Z|Y}$  is the Non-KSR operation and  $\hat{\mathcal{U}}_{Y|X}$  is the Mb-KSR operation in the setting given above. Then,

$$\|m_{Q_Z} - \hat{m}_{Q_Z}\|_{\mathcal{H}_Z}^2 = \xi^\top E \xi - 2w^\top \mathbf{m}_{Q_Z} + w^\top G_Z w, \quad (30)$$

where  $w = (G_Y + n\epsilon I_n)^{-1} G_{Y|\tilde{X}} \gamma$ .

*Proof* The derivation is similar to Proposition A12 and thus it is omitted.

## B Elliptical Distributions

This appendix presents definitions of elliptical (elliptically contoured) distributions. We check the condition of the characteristic property for positive-definite kernels generated by elliptical densities. We begin by reviewing spherical (spherically contoured) distributions.

**Definition B1** (McNeil et al, 2005, Definition 3.18, p. 89) A random vector  $X = (X_1, \dots, X_d)^\top$  on  $\mathbb{R}^d$  has a spherical distribution if it is distributionally invariant under rotations, i.e., it holds for every orthogonal matrix  $O \in \mathbb{R}^{d \times d}$  (matrix such that  $OO^\top = O^\top O = I_d$ ),

$$OX \stackrel{d}{=} X. \quad (31)$$

**Theorem B2** (McNeil et al, 2005, Theorem 3.19, p. 89) A random vector  $X$  on  $\mathbb{R}^d$  is spherical if and only if there a scalar variable function  $\phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{C}$  exists for the characteristic function  $\mathbb{E}_X[e^{\sqrt{-1}\theta^\top X}]$  such that

$$\mathbb{E}_X[e^{\sqrt{-1}\theta^\top X}] = \phi(\|\theta\|^2) = \phi(\theta^\top \theta) = \phi\left(\sum_{i=1}^d \theta_i^2\right), \quad \forall \theta \in \mathbb{R}^d. \quad (32)$$

A scalar variable function  $\phi$  fully describes a spherical random vector.  $\phi$  is called the *characteristic generator*. If  $X$  is spherical with  $\phi$ , we denote it as  $X \sim SC(\phi)$ . The normal distribution  $N(\mu, \Sigma)$  with  $\mu = 0$  and  $\Sigma = I_d$  is spherical and its characteristic generator is  $\phi(r) = e^{-\frac{1}{2}r}$ . If a spherical distribution has density, it is given by the form:

$$g(\|x\|^2) = g(x^\top x) = g\left(\sum_{i=1}^d x_i^2\right), \quad x \in \mathbb{R}^d, \quad (33)$$

with a scalar variable function  $g : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ .  $g$  is called the *density generator*.

An elliptical distribution is defined by an affine transformation of a spherical distribution.

**Definition B3** (McNeil et al, 2005, Definition 3.26, p. 93) A random vector  $X$  on  $\mathbb{R}^d$  has an elliptical distribution if  $X \stackrel{d}{=} \mu + AY$ , where  $Y \sim SC(\phi)$ ,  $A \in \mathbb{R}^{d \times m}$  is a matrix, and  $\mu \in \mathbb{R}^d$  is a vector.

The characteristic function of an elliptical distribution is given by  $\mathbb{E}_X[e^{\sqrt{-1}\theta^\top X}] = e^{\sqrt{-1}\theta^\top \mu} \phi(\theta^\top \Sigma \theta)$  with  $\Sigma = AA^\top$ . If  $X$  is an elliptical random vector with  $\mu$ ,  $\Sigma$ , and  $\phi$ , then we denote it as  $X \sim EC(\mu, \Sigma, \phi)$ <sup>5</sup>. If  $\Sigma$  is positive-definite and  $g$  is the density generator of  $Y$ , then  $X$  has the density

$$d_{EC}(x; \mu, \Sigma, \phi) = \frac{1}{|\Sigma|^{1/2}} g(\|x - \mu\|_{\Sigma^{-1}}^2) = \frac{1}{|\Sigma|^{1/2}} g((x - \mu)^\top \Sigma^{-1} (x - \mu)). \quad (34)$$

The following condition is required for an elliptical kernel to be characteristic. This condition guarantees that a kernel mean uniquely specifies its probability distribution when the elliptical kernel is used, e.g., for the Mb-KSR.

**Proposition B4** Suppose that  $d_{EC}(x; 0, \Sigma, \phi)$ ,  $x \in \mathbb{R}^d$  is a bounded continuous elliptical density on  $\mathbb{R}^d$ . A function  $k_{\Sigma, \phi}(x, y) = d_{EC}(x - y; 0, \Sigma, \phi)$ ,  $x, y \in \mathbb{R}^d$  is a positive-definite kernel if and only if  $\phi$  is a nonnegative function  $\phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ . The elliptical kernel  $k_{\Sigma, \phi}(x, y)$  is characteristic if and only if  $\phi$  is a positive function  $\phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{> 0}$ .

*Proof* The symmetric function  $k_{\Sigma, \phi}(x, y)$ ,  $x, y \in \mathbb{R}^d$  is positive-definite if and only if the Fourier transform of  $d_{EC}(x; 0, \Sigma, \phi)$ ,  $x \in \mathbb{R}^d$  is a finite nonnegative measure on  $\mathbb{R}^d$  (Bochner's Theorem). The elliptical kernel  $k_{\Sigma, \phi}(x, y)$ ,  $x, y \in \mathbb{R}^d$  is characteristic if and only if the Fourier transform of  $d_{EC}(x; 0, \Sigma, \phi)$ ,  $x \in \mathbb{R}^d$  has the entire support  $\mathbb{R}^d$  (Sriperumbudur et al, 2010, Theorem 9).

**Acknowledgements** This research was partly supported by JSPS KAKENHI (B) 22300098, MEXT Grant-in-Aid for Scientific Research on Innovative Areas 25120012, and JSPS Wakate (B) 26870821.

<sup>5</sup> The representation  $EC(\mu, \Sigma, \phi)$  is unique up to a positive constant for  $\Sigma$  and  $\phi$ . For example, the normal distribution  $N(\mu, \Sigma)$  with the characteristic generator  $\phi(r) = e^{-\frac{1}{2}r}$  has multiple expressions  $EC(\mu, c\Sigma, \phi(\cdot/c))$  for any  $c > 0$ .

## References

- Boots B, Gordon G, Gretton A (2013) Hilbert Space Embeddings of Predictive State Representations. In: UAI, pp 92–101
- Fukumizu K, Leng C (2012) Gradient-based kernel method for feature extraction and variable selection. In: NIPS, pp 2123–2131
- Fukumizu K, Bach FR, Jordan MI (2004) Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research* 5:73–99
- Fukumizu K, Gretton A, Sun X, Schölkopf B (2008) Kernel Measures of Conditional Dependence. In: NIPS, pp 489–496
- Fukumizu K, Song L, Gretton A (2011) Kernel Bayes' Rule. In: NIPS, pp 1737–1745
- Fukumizu K, Song L, Gretton A (2013) Kernel bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research* pp 3753–3783
- Gretton A, Györfi L (2010) Consistent Nonparametric Tests of Independence. *Journal of Machine Learning Research* 11:1391–1423
- Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola AJ (2012) A Kernel Two-Sample Test. *Journal of Machine Learning Research* 13:723–773
- Grünewälder S, Lever G, Baldassarre L, Pontil M, Gretton A (2012) Modelling transition dynamics in MDPs with RKHS embeddings. In: ICML, pp 535–542
- Kanagawa M, Fukumizu K (2014) Recovering Distributions from Gaussian RKHS Embeddings. In: AISTATS, pp 457–465
- Kanagawa M, Nishiyama Y, Gretton A, Fukumizu K (2014) Monte Carlo Filtering Using Kernel Embedding of Distributions. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, pp 1897–1903
- Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp 2169–2178
- Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110
- McCalman L, O'Callaghan S, Ramos F (2013) Multi-modal estimation with kernel embeddings for learning motion models. In: IEEE International Conference on Robots and Automation (ICRA)
- McNeil A, Frey R, Embrechts P (2005) Quantitative Risk Management. Princeton University Press
- Mika S, Schölkopf B, Smola A, Müller K, Scholz M, Rätsch G (1999) Kernel PCA and denoising in feature spaces. In: NIPS, pp 536–542
- Muandet K, Fukumizu K, Dinuzzo F, Schölkopf B (2012) Learning from Distributions via Support Measure Machines. In: NIPS, pp 10–18
- Nishiyama Y, Fukumizu K (2014) Characteristic Kernels and Infinitely Divisible Distributions. arXiv:14037304
- Nishiyama Y, Boularias A, Gretton A, Fukumizu K (2012) Hilbert Space Embeddings of POMDPs. In: UAI, pp 644–653
- Pronobis A, Caputo B (2009) COLD: COsy Localization Database. *The International Journal of Robotics Research (IJRR)* 28(5):588–594
- Rawlik K, Toussaint M, Vijayakumar S (2013) Path Integral Control by Reproducing Kernel Hilbert Space Embedding. *Proc 23rd Int Joint Conference on Artificial Intelligence (IJCAI)*
- Schölkopf B, Smola A (2002) Learning with Kernels. MIT Press, Cambridge
- Smola A, Gretton A, Song L, Schölkopf B (2007) A Hilbert space embedding for distributions. In: International Conference on Algorithmic Learning Theory (ALT), pp 13–31
- Song L, Zhang X, Smola A, Gretton A, Schölkopf B (2008) Tailoring Density Estimation via Reproducing Kernel Moment Matching. *ICML* pp 992–999
- Song L, Huang J, Smola A, Fukumizu K (2009) Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In: ICML, pp 961–968
- Song L, Boots B, Siddiqi SM, Gordon GJ, Smola AJ (2010) Hilbert Space Embeddings of Hidden Markov Models. In: ICML, pp 991–998
- Song L, Gretton A, Bickson D, Low Y, Guestrin C (2011) Kernel Belief Propagation. *Journal of Machine Learning Research - Proceedings Track* 15:707–715
- Song L, Fukumizu K, Gretton A (2013) Kernel embedding of conditional distributions. *IEEE Signal Processing Magazine* 30(4):98–111
- Sriperumbudur B, Gretton A, Fukumizu K, Lanckriet G, Schölkopf B (2010) Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*

---

11:1517–1561

Steinwart I, Christmann A (2008) Support Vector Machines. Information Science and Statistics. Springer

Thrun S, Burgard W, Fox D (2005) Probabilistic Robotics. MIT Press, Cambridge, MA

Vlassis N, Terwijn B, Kröwe B (2002) Auxiliary particle filter robot localization from high-dimensional sensor observations. In: ICRA, pp 7–12